Stochastic optimization in machine learning

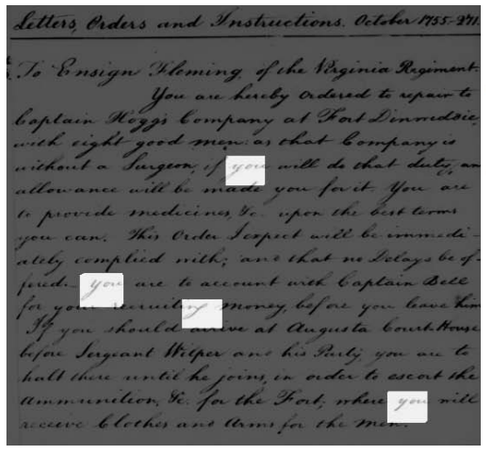
Jalal Fadili
Normandie Université-ENSICAEN, CNRS

School CIMPA
26-30 May 2025

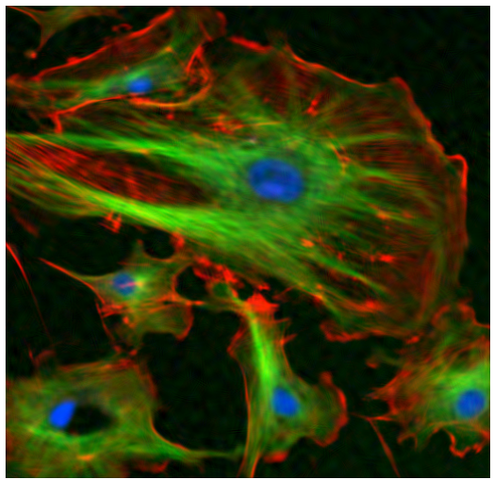
Motivations

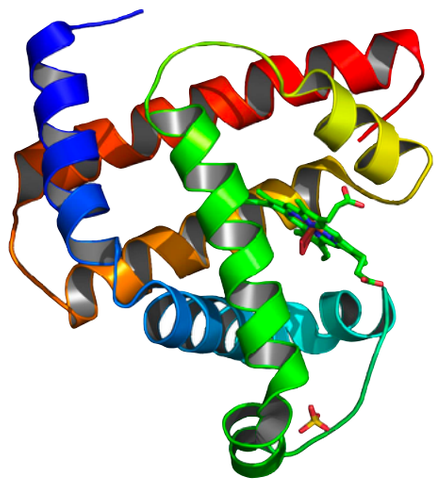
- Optimization and statistics are the cornerstone of modern data science.

Language models, generative models

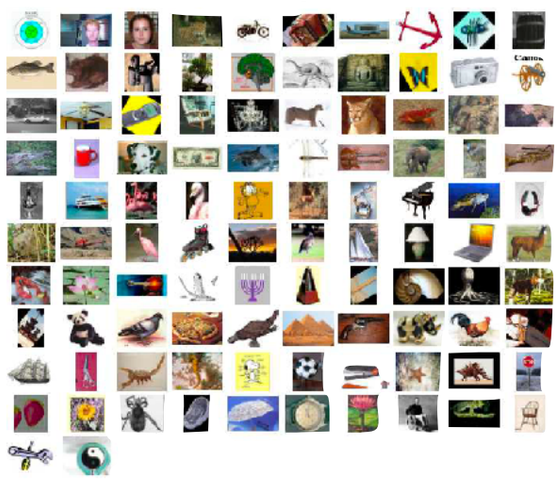


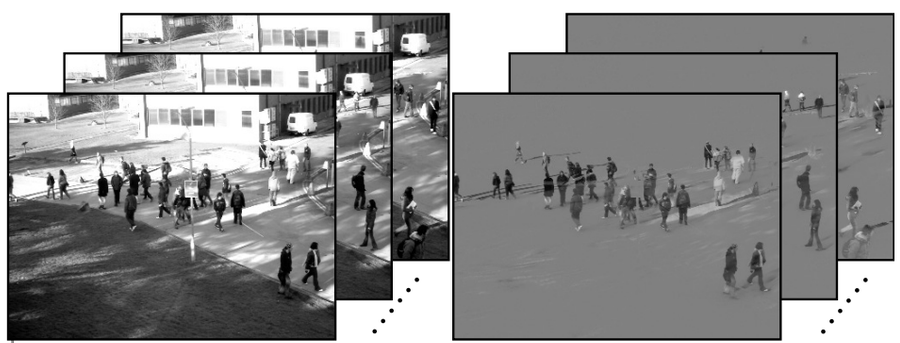
Bio



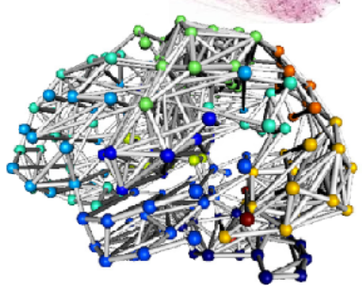


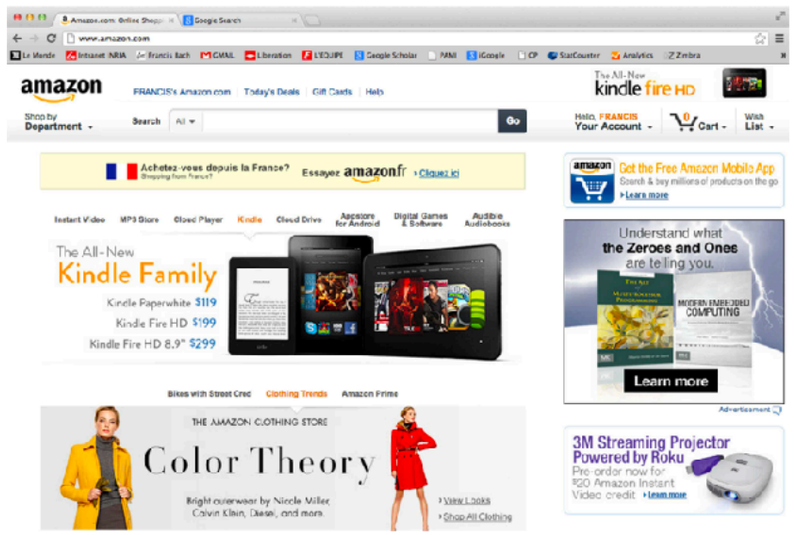
Computer vision & graphics





Networks





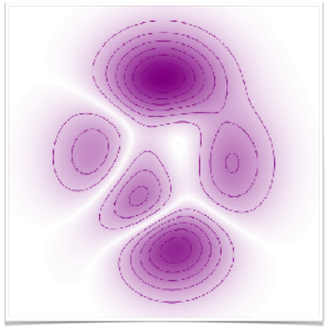
Search engines
Recommendation systems

Motivations

- Optimization and statistics are the cornerstone of modern data science.
- Learning from large “scale” data:
  - n observations in dimension d.
  - Both large.
- At the interface of :
  - mathematics: optimization, statistics, probability.
  - computer science.
- Goals :
  - Develop algorithms.
  - Their theoretical guarantees.
  - Efficient implementations.

Supervised machine learning

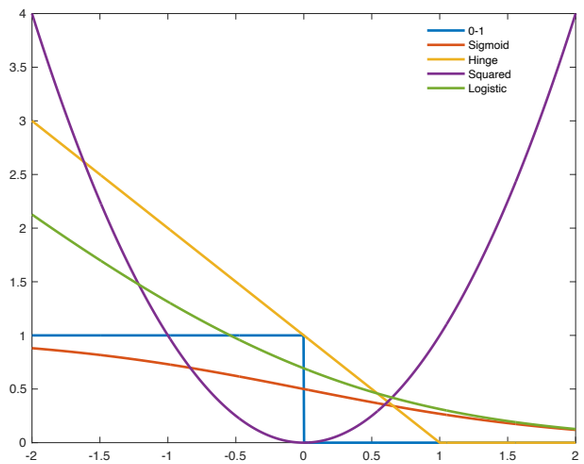
- Data : n observations $(u_i, v_i) \in \mathcal{U} \times \mathcal{V}$, $i = 1, \ldots, n$, i.i.d. drawn from some probability measure on $\mathcal{U} \times \mathcal{V}$.



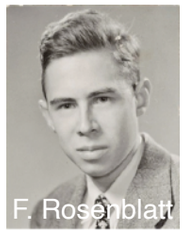
  - $\mathcal{U}$: space of inputs.
  - $\mathcal{V}$: space of outputs.
- Prediction as a linear function $x^\top \varphi(u)$ of features $\varphi(u)$, $\varphi : \mathcal{U} \to \mathbb{R}^d$ is measurable.
- Many supervised machine learning models boil down to solving the (regularized) empirical risk minimization problem

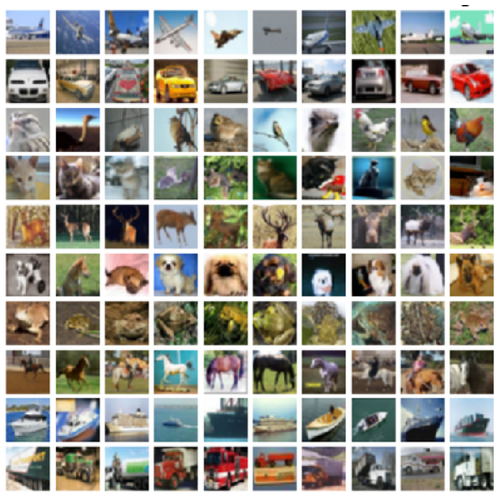
$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{n} \sum_{i=1}^{n} \ell(v_i, x^\top \varphi(u_i)) + \lambda R(x) \right\}.$$

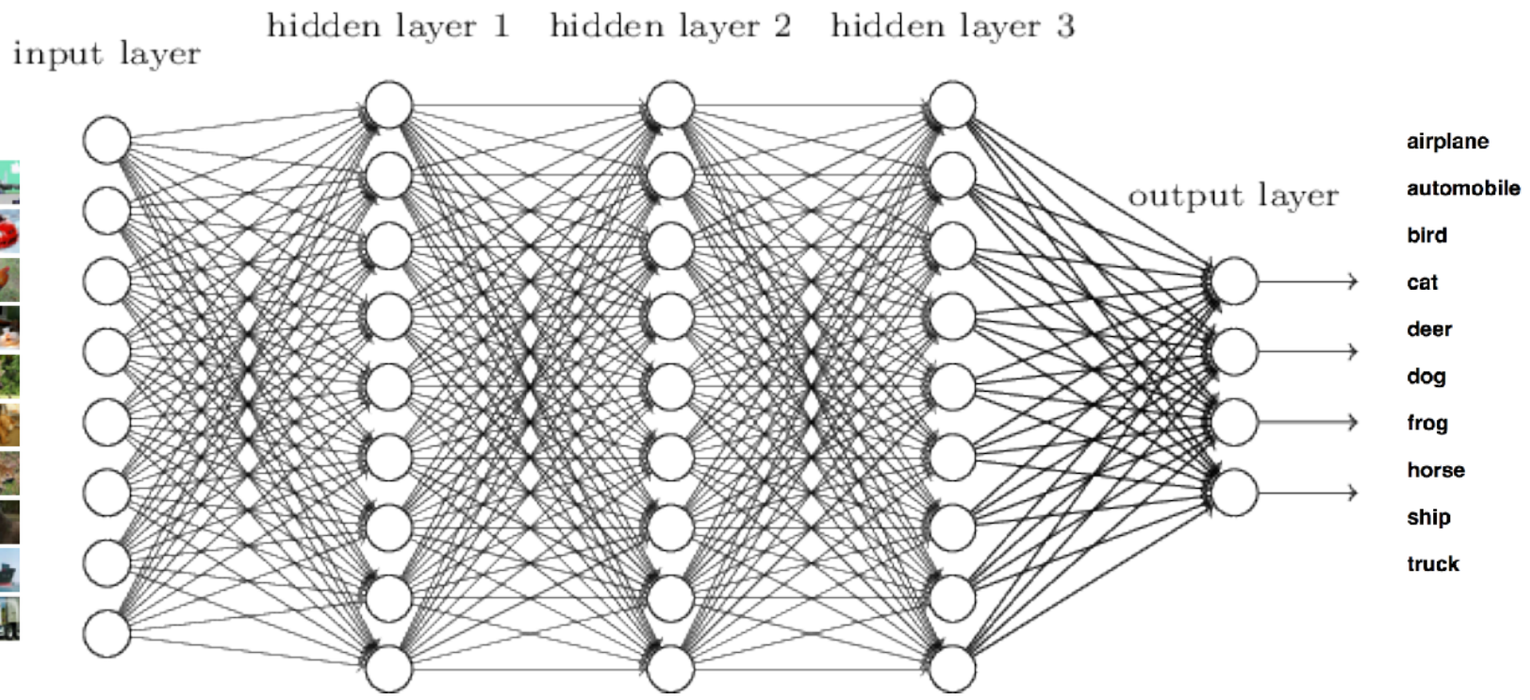
  - $\lambda \geq 0$: regularization parameter ;
  - $\ell : \mathbb{R} \times \mathcal{V} \to \mathbb{R}$: the loss function ;
  - $R : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ the regularization function.
  - The minimum is assumed to be attained.



Neural Networks (NNs)

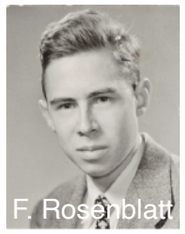


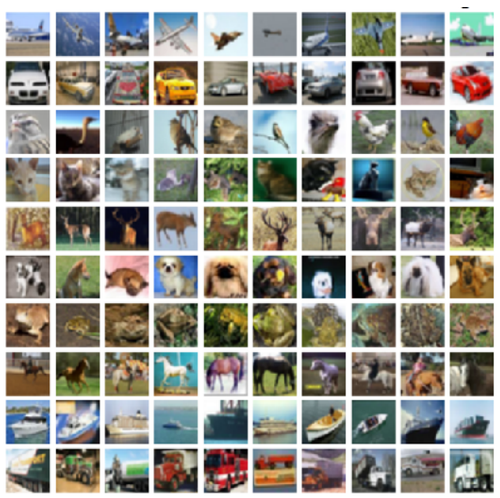


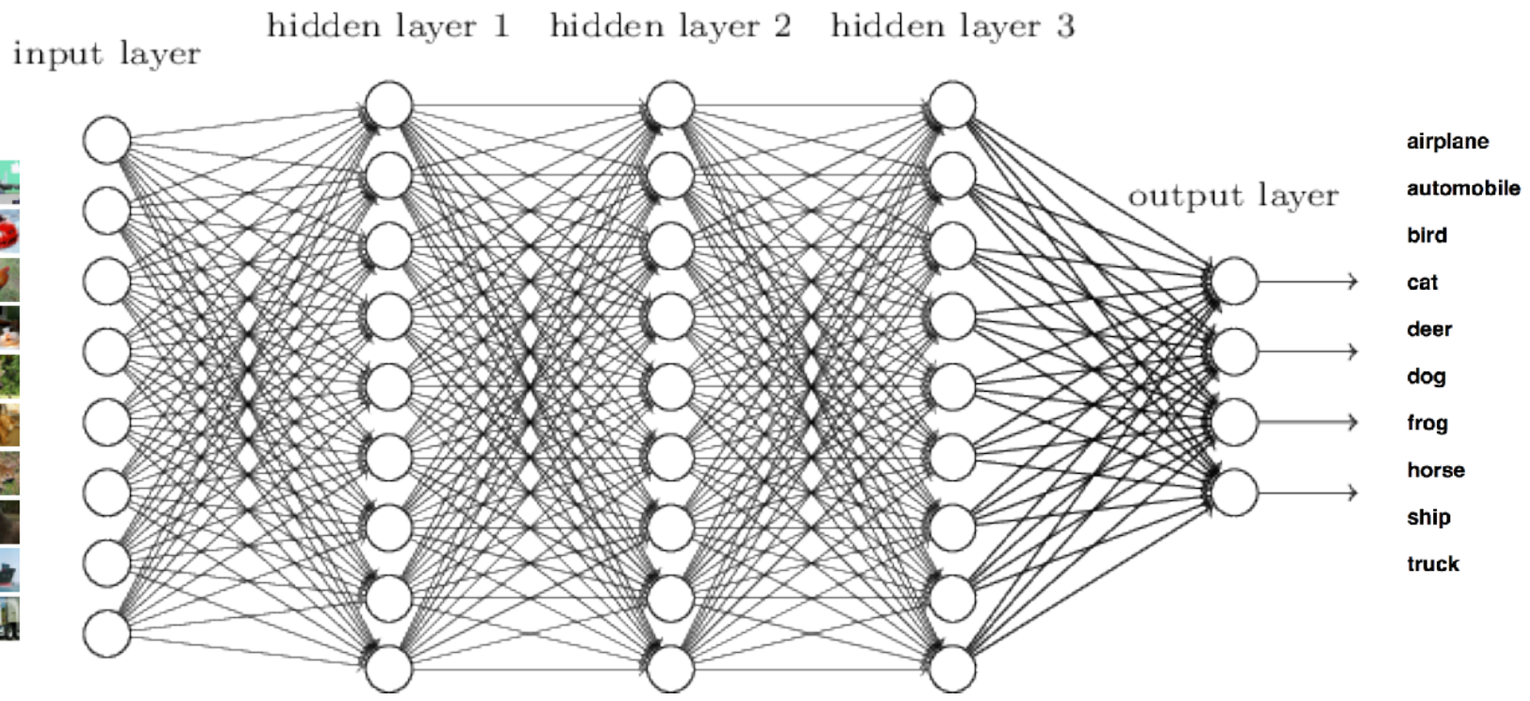


Neural Networks (NNs)

- NNs have a long history (50's).

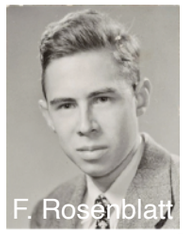




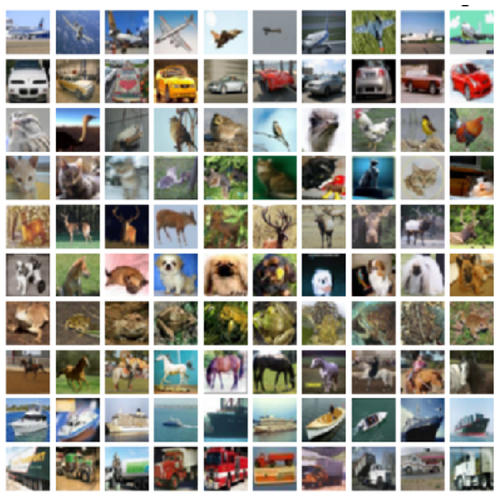


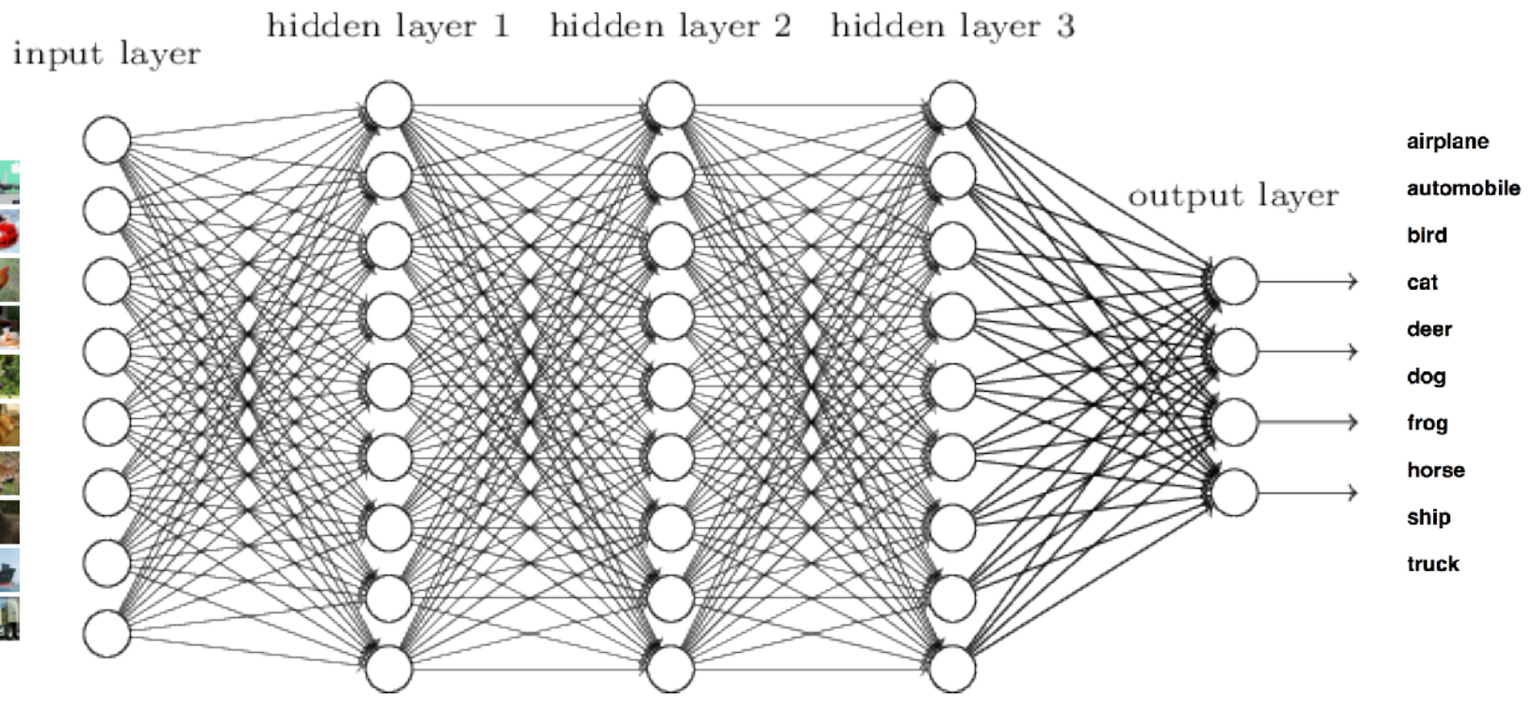
Neural Networks (NNs)

- NNs have a long history (50's).



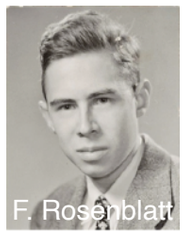
- For many people, deep learning is another name (confusion) for a set of algorithms that use a neural network as an architecture ….



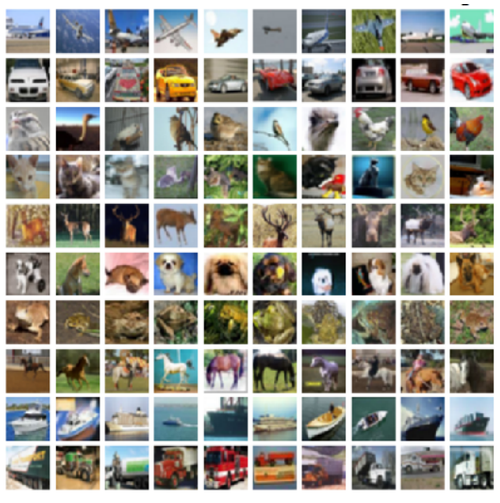


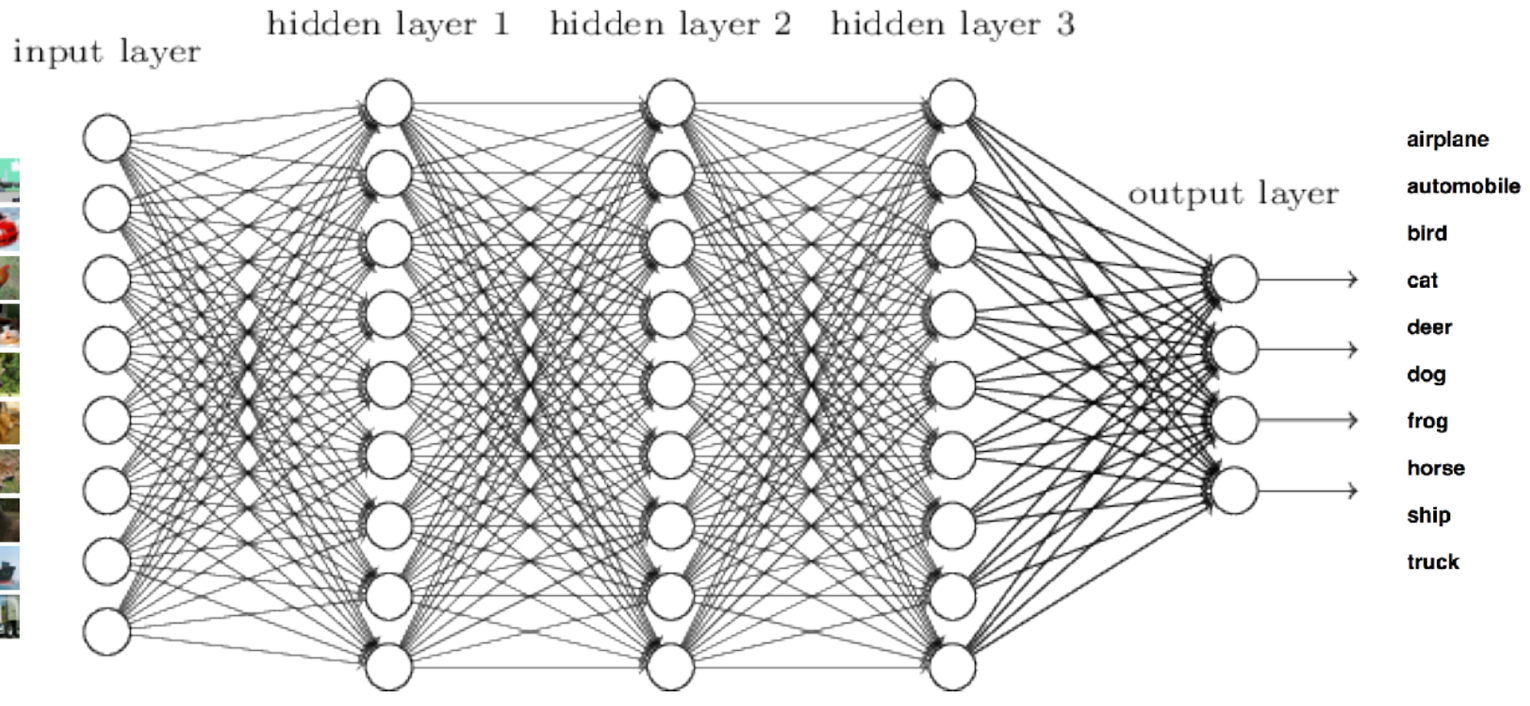
Neural Networks (NNs)

- NNs have a long history (50's).



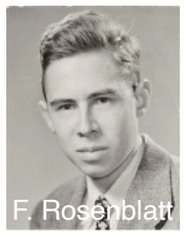
- For many people, deep learning is another name (confusion) for a set of algorithms that use a neural network as an architecture ….
- NNs gain tremendous success in recent years due to:
  - the availability of inexpensive, parallel hardware (GPUs, computer clusters),
  - massive amounts of data, and
  - the recent development of efficient optimization algorithms.



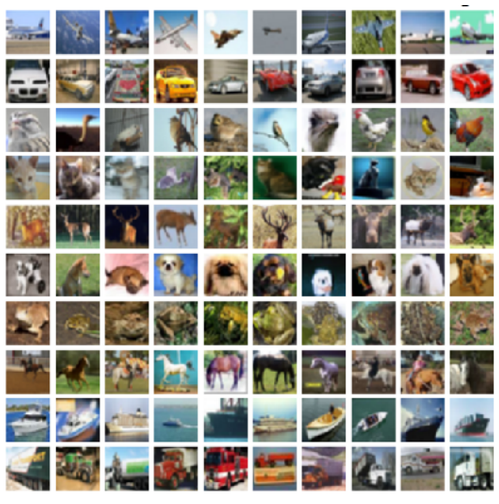


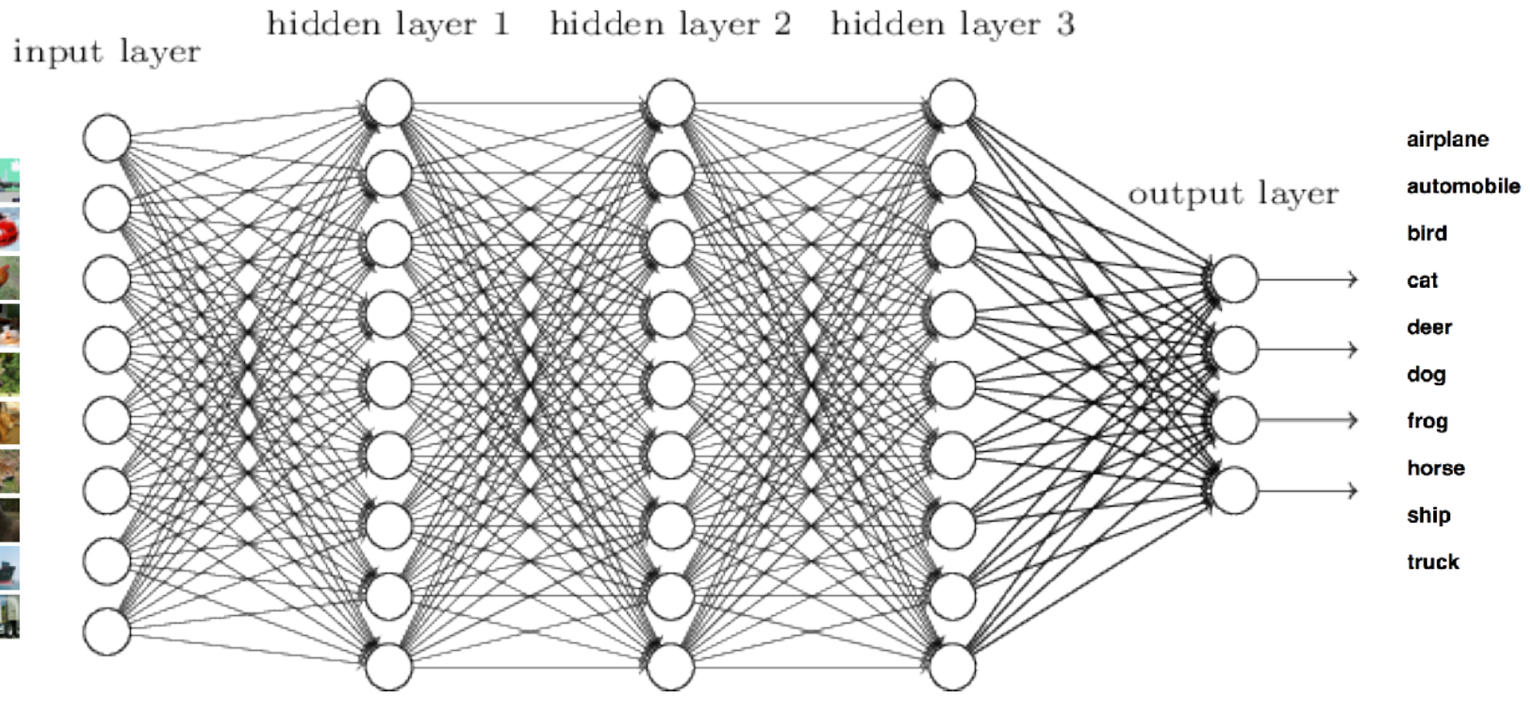
Neural Networks (NNs)

- NNs have a long history (50's).

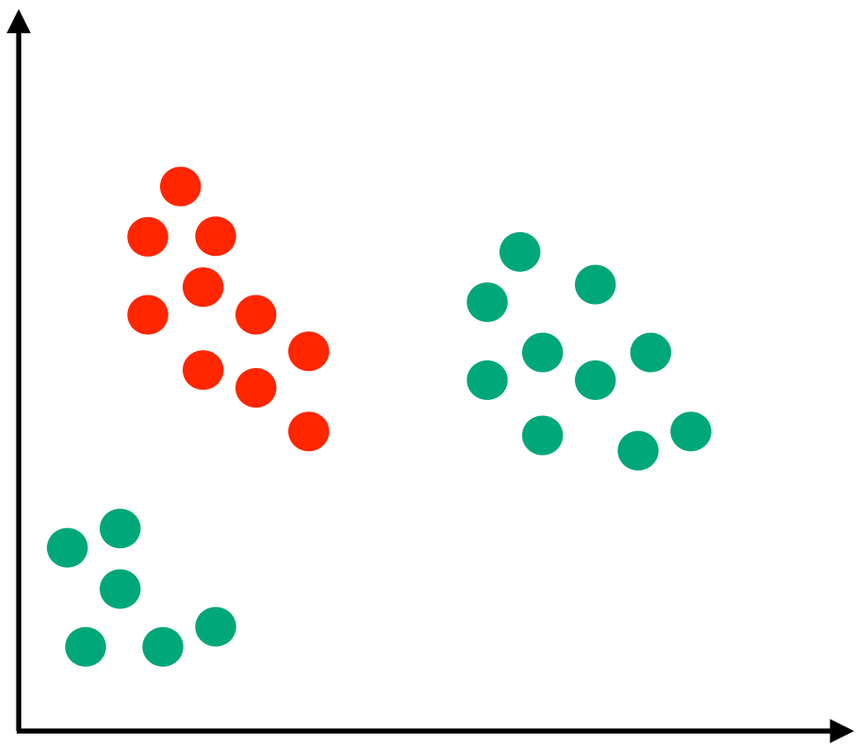


- For many people, deep learning is another name (confusion) for a set of algorithms that use a neural network as an architecture ….
- NNs gain tremendous success in recent years due to:
  - the availability of inexpensive, parallel hardware (GPUs, computer clusters),
  - massive amounts of data, and
  - the recent development of efficient optimization algorithms.
- A plethora of architectures (art of design), with millions or even hundreds of billions of neurons.



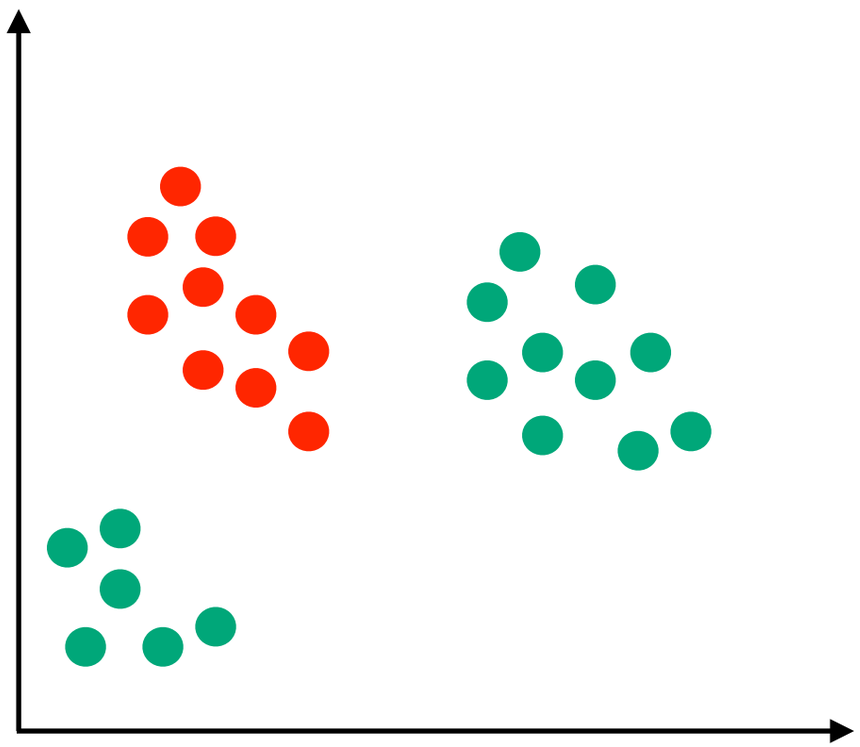


A glimpse of Neural Networks (NNs)



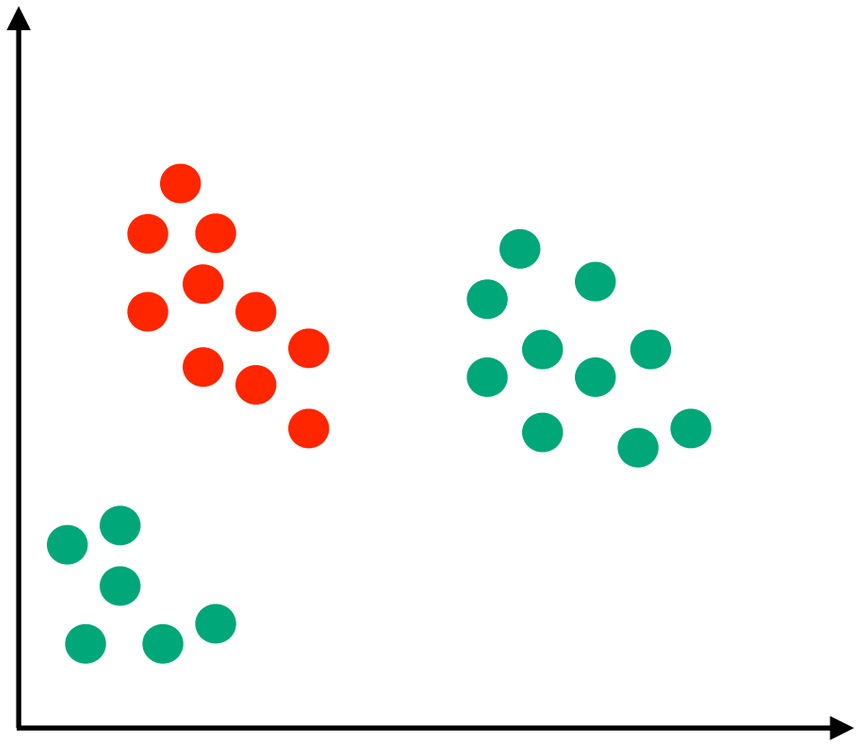
A glimpse of Neural Networks (NNs)

- Non-linearly separable classification problem.



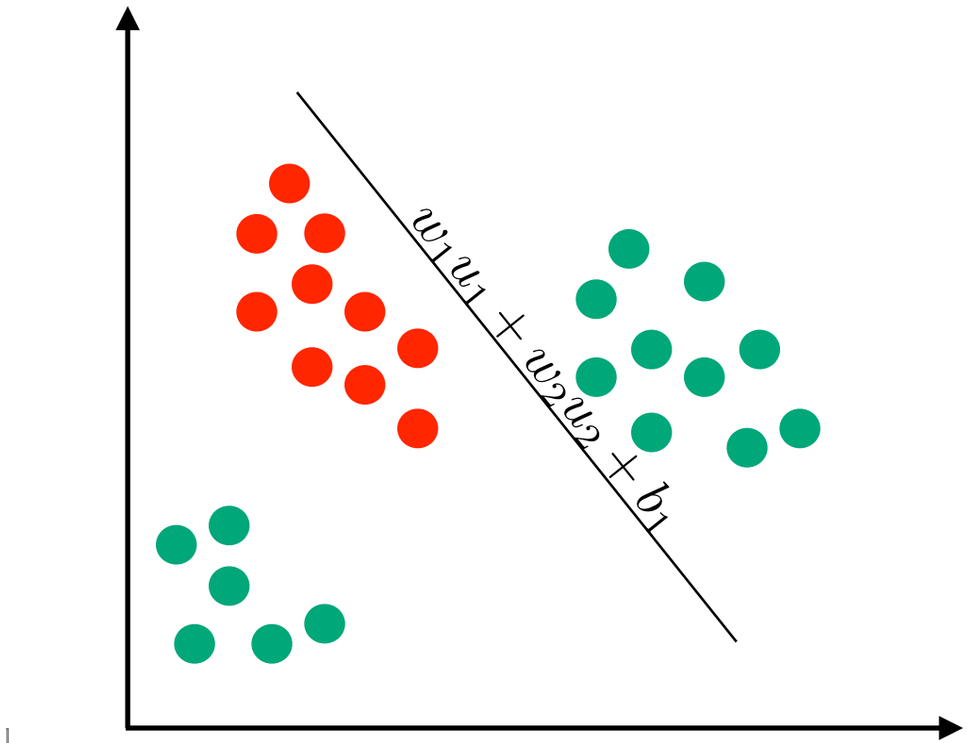
A glimpse of Neural Networks (NNs)

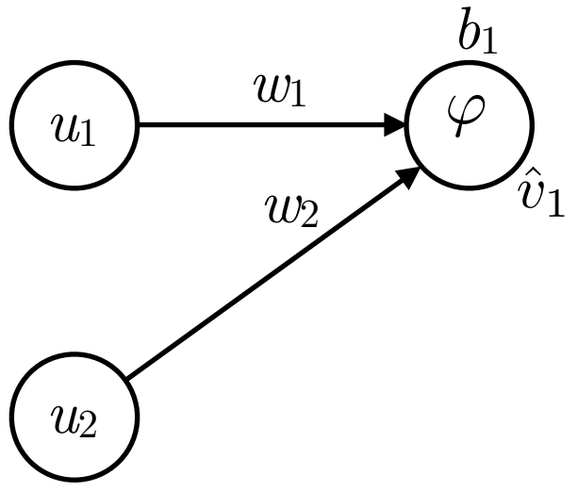
- Non-linearly separable classification problem.
- A heuristic approach this complex problem: decompose it into smaller problems that one can solve.



A glimpse of Neural Networks (NNs)

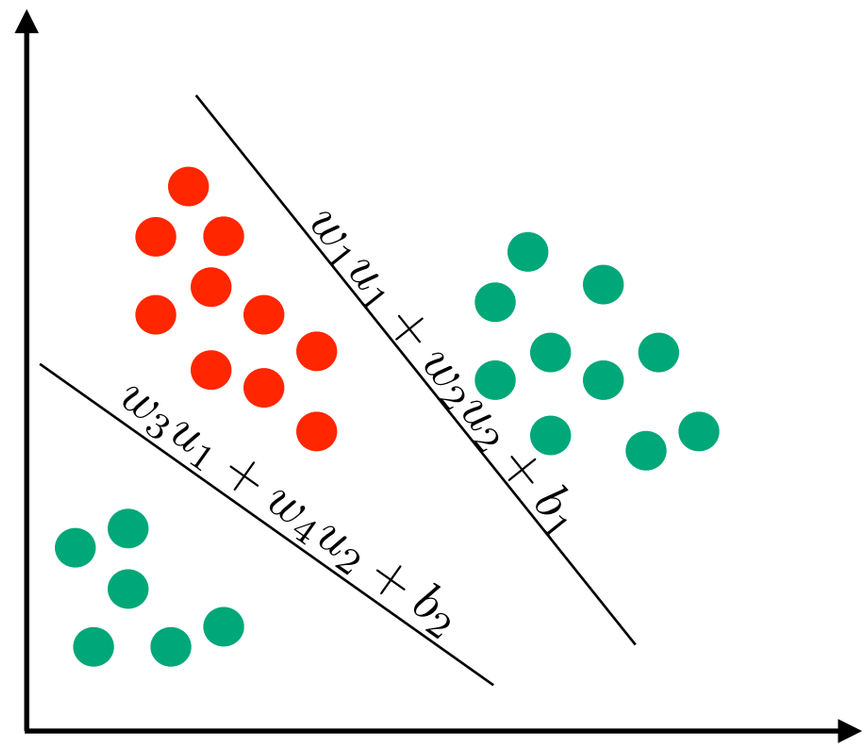
- Non-linearly separable classification problem.
- A heuristic approach this complex problem: decompose it into smaller problems that one can solve.
  - Throw away the “weird” examples from the bottom left corner ⇒ linearly separable problem.

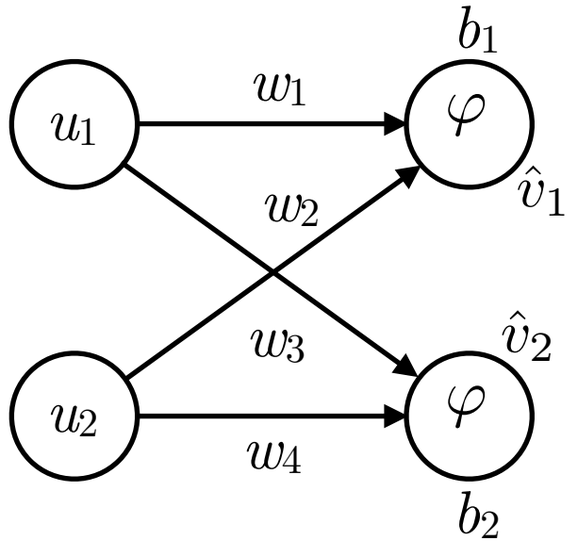




A glimpse of Neural Networks (NNs)

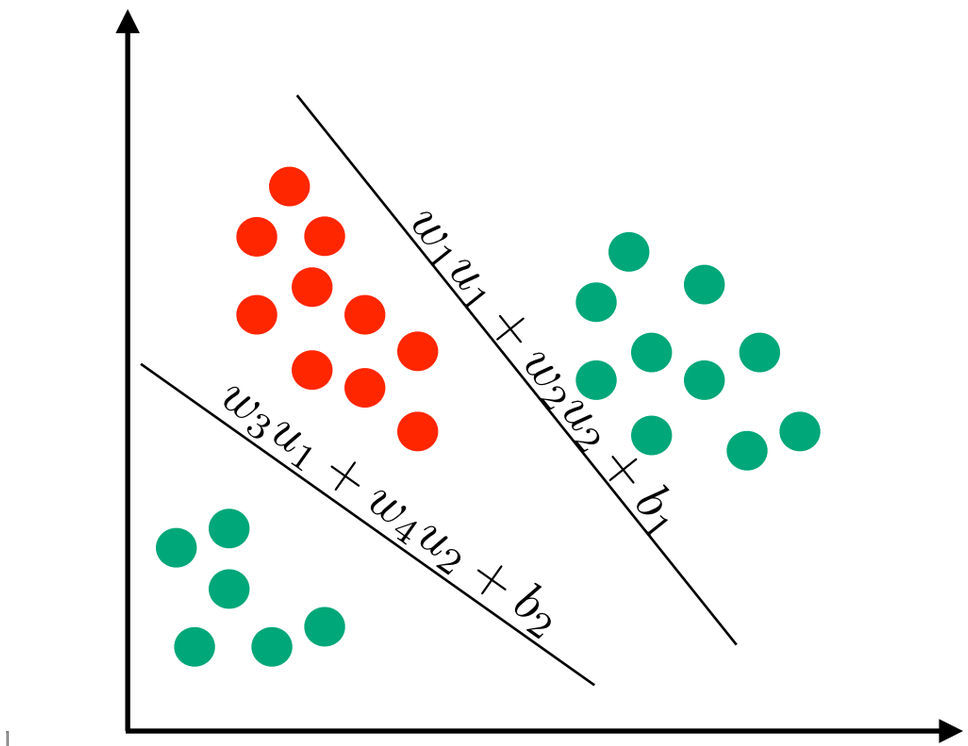
- Non-linearly separable classification problem.
- A heuristic approach this complex problem: decompose it into smaller problems that one can solve.
  - Throw away the “weird” examples from the bottom left corner ⇒ linearly separable problem.
  - Throw away the “weird” examples from the top right corner ⇒ linearly separable problem.

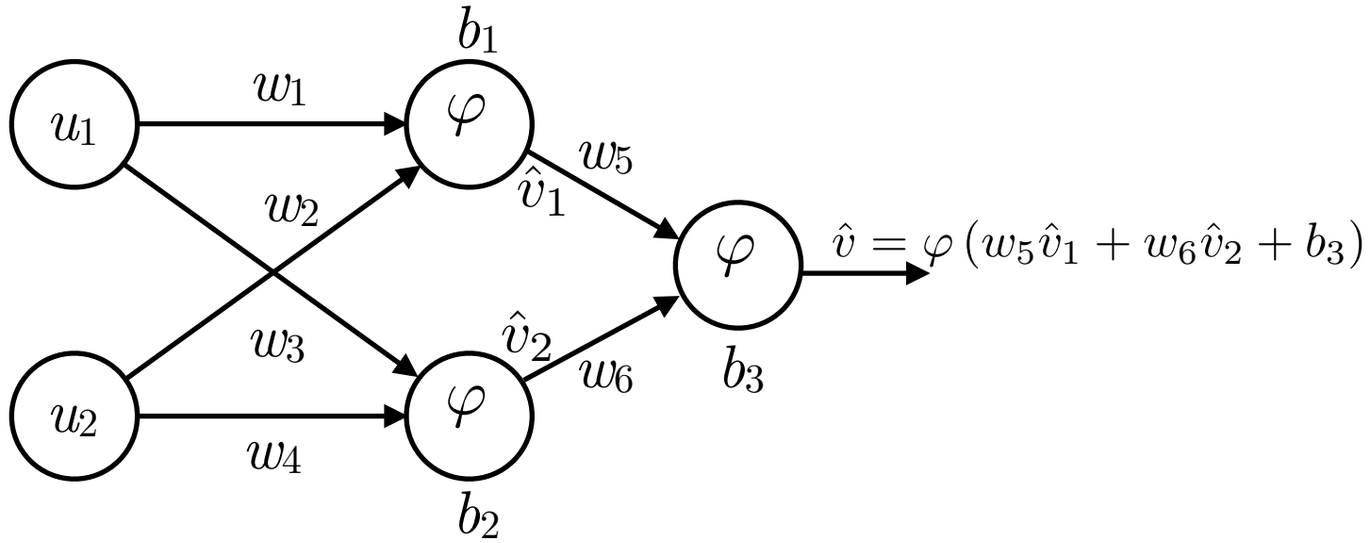




A glimpse of Neural Networks (NNs)

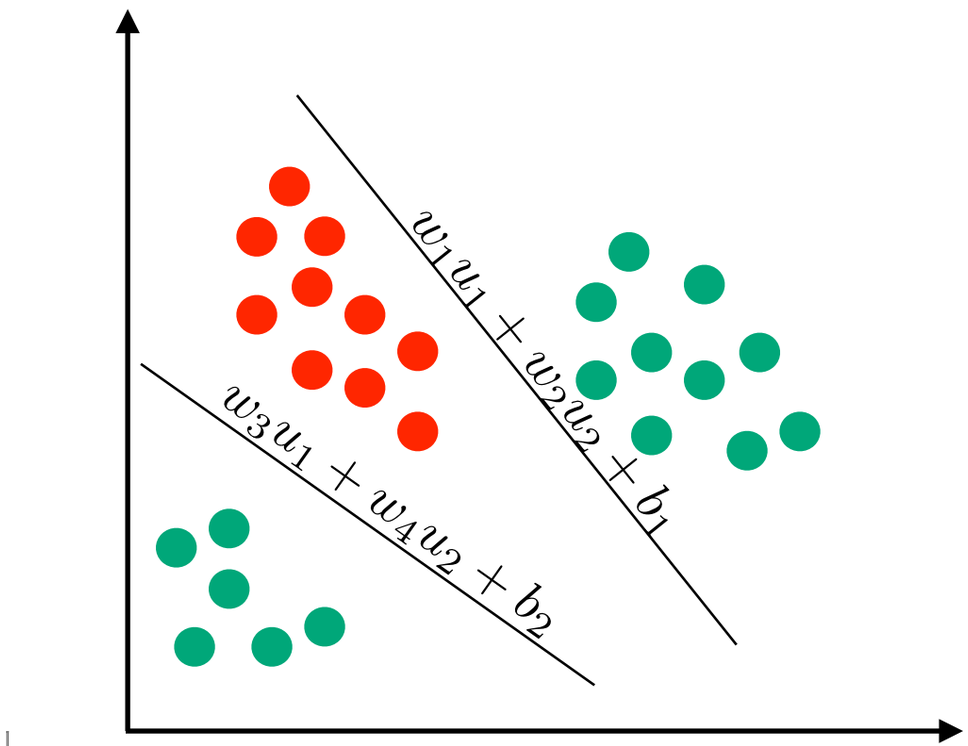
- Non-linearly separable classification problem.
- A heuristic approach this complex problem: decompose it into smaller problems that one can solve.
  - Throw away the “weird” examples from the bottom left corner ⇒ linearly separable problem.
  - Throw away the “weird” examples from the top right corner ⇒ linearly separable problem.
  - Combine these two decision functions into one final decision function.

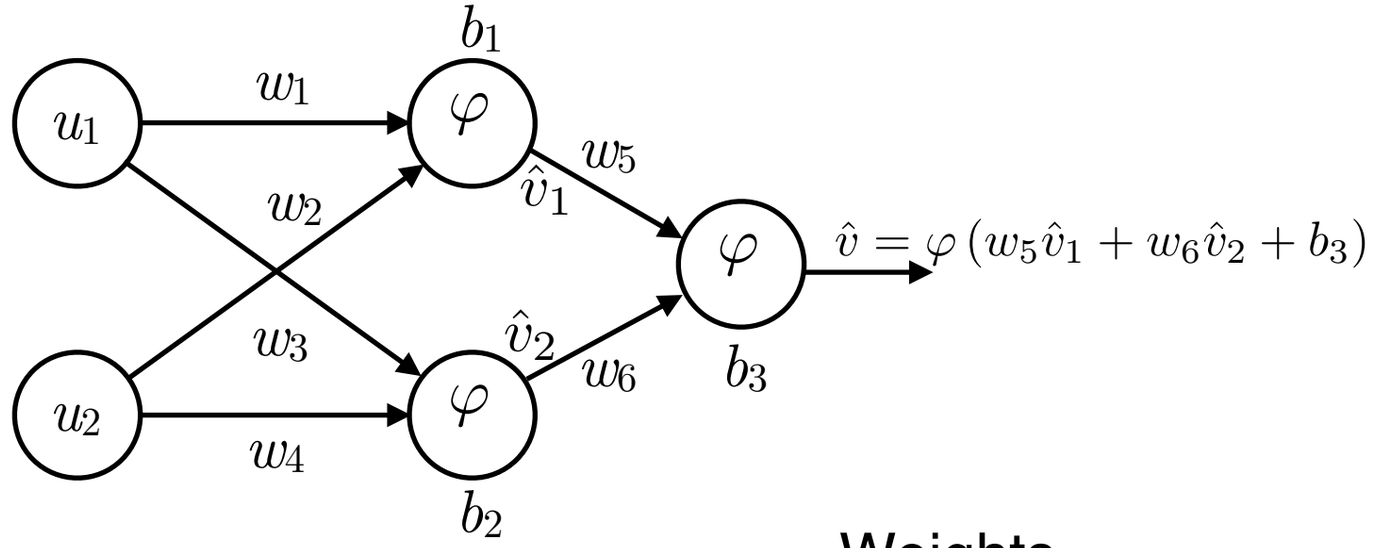




A glimpse of Neural Networks (NNs)

- Non-linearly separable classification problem.
- A heuristic approach this complex problem: decompose it into smaller problems that one can solve.
  - Throw away the “weird” examples from the bottom left corner ⇒ linearly separable problem.
  - Throw away the “weird” examples from the top right corner ⇒ linearly separable problem.
  - Combine these two decision functions into one final decision function.



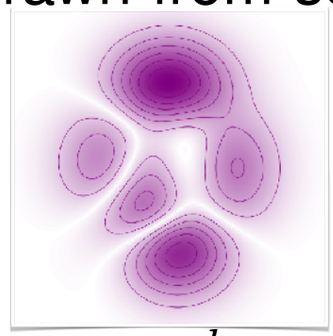


$$\min_{w\in\mathbb{R}^6, b\in\mathbb{R}^3} \frac{1}{n}\sum_{i=1}^{n} \ell(v_i, \hat{v}_i(w, b)).$$

Weights ↓ … ↑ Biases

Empirical vs population risk

- Data : n observations $(u_i, v_i) \in \mathcal{U} \times \mathcal{V}$, $i = 1, \ldots, n$, i.i.d. drawn from some probability measure on (the measurable space) $\mathcal{U} \times \mathcal{V}$.



  - $\mathcal{U}$: space of inputs.
  - $\mathcal{V}$: space of outputs.
- Prediction as a linear function $x^\top \varphi(u)$ of features $\varphi(u)$, $\varphi : \mathcal{U} \to \mathbb{R}^d$ is measurable.
- Many supervised machine learning models boil down to solving the (regularized) empirical risk minimization problem

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(v_i, x^\top \varphi(u_i)) \quad \text{s.t.} \quad R(x) \leq c.$$

  - Empirical risk $\Rightarrow$ training cost : $\widehat{L}(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \ell(v_i, x^\top \varphi(u_i))$.
  - Population/expected risk $\Rightarrow$ testing cost : $L(x) \stackrel{\text{def}}{=} \mathbb{E}_{u,v}(\ell(v, x^\top \varphi(u)))$.
- Two main questions :
  - Solve the empirical risk minimization problem (**optimization**).
  - Analyze its properties, and in particular its relation to the population risk minimization (**generalization**).

Generalization and uniform convergence

- Data : n observations $(u_i, v_i) \in \mathcal{U} \times \mathcal{V}$, $i = 1, \ldots, n$, i.i.d. drawn from some probability measure $\mathbb{P}$ on (the measurable space) $\mathcal{U} \times \mathcal{V}$.
- Prediction $g : \mathcal{U} \to \mathcal{V}$ (special case : $g(u) = x^\top \varphi(u)$).
- g is chosen from a set of functions $\mathcal{G} \subseteq \mathcal{V}^{\mathcal{U}}$ (e.g., the set of linear predictors $x^\top \varphi(\cdot)$, set of functions computed by a NN, etc.).
- Empirical risk : $\widehat{L}(g) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \ell(v_i, g(u_i))$.
- Expected risk : $L(g) \stackrel{\text{def}}{=} \mathbb{E}_{u,v}(\ell(v, g(u)))$.
- $L(g)$ is random as it is chosen based on random training data.

Generalization and uniform convergence

- Data : n observations $(u_i, v_i) \in \mathcal{U} \times \mathcal{V}$, $i = 1, \ldots, n$, i.i.d. drawn from some probability measure $\mathbb{P}$ on (the measurable space) $\mathcal{U} \times \mathcal{V}$.
- Prediction $g : \mathcal{U} \to \mathcal{V}$ (special case : $g(u) = x^\top \varphi(u)$).
- g is chosen from a set of functions $\mathcal{G} \subseteq \mathcal{V}^{\mathcal{U}}$ (e.g., the set of linear predictors $x^\top \varphi(\cdot)$, set of functions computed by a NN, etc.).
- Empirical risk : $\widehat{L}(g) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \ell(v_i, g(u_i))$.
- Expected risk : $L(g) \stackrel{\text{def}}{=} \mathbb{E}_{u,v}(\ell(v, g(u)))$.
- $L(g)$ is random as it is chosen based on random training data.
- Naturally choose the ERM predictor

$$g^\star_{\text{erm}} \in \operatorname*{Argmin}_{g \in \mathcal{G}} \widehat{L}(g).$$

- Fundamental question : how well $g^\star_{\text{erm}}$ predicts the relationship between all pairs $(u, v) \sim \mathbb{P}$ in the sense that its population risk is close to the one of the best possible predictor, i.e. the excess risk

$$L(g^\star_{\text{erm}}) - \inf_g L(g) \quad \text{is small.}$$

Generalization and uniform convergence

$$L(g^\star_{\mathrm{erm}}) - \inf_g L(g) = \left(L(g^\star_{\mathrm{erm}}) - \inf_{g \in \mathcal{G}} L(g) \right) + \left(\inf_{g \in \mathcal{G}} L(g) - \inf_g L(g) \right)$$

Generalization and uniform convergence

$$L(g^\star_{\mathrm{erm}}) - \inf_g L(g) = \underbrace{\left(L(g^\star_{\mathrm{erm}}) - \inf_{g\in\mathcal{G}} L(g)\right)}_{\text{Estimation error}} + \underbrace{\left(\inf_{g\in\mathcal{G}} L(g) - \inf_g L(g)\right)}_{\substack{\text{Approximation error}\\ \text{small if } \mathcal{G} \text{ rich enough}}}$$

Generalization and uniform convergence

$$L(g^\star_{\mathrm{erm}}) - \inf_g L(g) = \underbrace{\left(L(g^\star_{\mathrm{erm}}) - \inf_{g\in\mathcal{G}} L(g)\right)}_{\text{Estimation error}} + \underbrace{\left(\inf_{g\in\mathcal{G}} L(g) - \inf_g L(g)\right)}_{\substack{\text{Approximation error}\\ \text{small if } \mathcal{G} \text{ rich enough}}}$$

$g^\star \in \operatorname*{Argmin}_{g\in\mathcal{G}} L(g) \neq \emptyset$

$$L(g^\star_{\mathrm{erm}}) - L(g^\star) = \left(L(g^\star_{\mathrm{erm}}) - \widehat{L}(g^\star_{\mathrm{erm}})\right) + \left(\widehat{L}(g^\star_{\mathrm{erm}}) - \widehat{L}(g^\star)\right) + \left(\widehat{L}(g^\star) - L(g^\star)\right)$$

Generalization and uniform convergence

$$L(g^\star_{\mathrm{erm}}) - \inf_g L(g) = \underbrace{\left(L(g^\star_{\mathrm{erm}}) - \inf_{g\in\mathcal{G}} L(g)\right)}_{\text{Estimation error}} + \underbrace{\left(\inf_{g\in\mathcal{G}} L(g) - \inf_g L(g)\right)}_{\substack{\text{Approximation error}\\ \text{small if } \mathcal{G} \text{ rich enough}}}$$

$g^\star \in \operatorname*{Argmin}_{g\in\mathcal{G}} L(g) \neq \emptyset$

$$L(g^\star_{\mathrm{erm}}) - L(g^\star) = \left(L(g^\star_{\mathrm{erm}}) - \widehat{L}(g^\star_{\mathrm{erm}})\right) + \overbrace{\left(\widehat{L}(g^\star_{\mathrm{erm}}) - \widehat{L}(g^\star)\right)}^{\leq 0 \text{ by optimality}} + \left(\widehat{L}(g^\star) - L(g^\star)\right)$$

Generalization and uniform convergence

$$L(g^\star_{\mathrm{erm}}) - \inf_g L(g) = \underbrace{\left(L(g^\star_{\mathrm{erm}}) - \inf_{g\in\mathcal{G}} L(g)\right)}_{\text{Estimation error}} + \underbrace{\left(\inf_{g\in\mathcal{G}} L(g) - \inf_g L(g)\right)}_{\substack{\text{Approximation error}\\ \text{small if } \mathcal{G} \text{ rich enough}}}$$

$g^\star \in \operatorname*{Argmin}_{g\in\mathcal{G}} L(g) \neq \emptyset$

$$L(g^\star_{\mathrm{erm}}) - L(g^\star) = \left(L(g^\star_{\mathrm{erm}}) - \widehat{L}(g^\star_{\mathrm{erm}})\right) + \overbrace{\left(\widehat{L}(g^\star_{\mathrm{erm}}) - \widehat{L}(g^\star)\right)}^{\leq 0 \text{ by optimality}} + \overbrace{\left(\widehat{L}(g^\star) - L(g^\star)\right)}^{\to 0 \text{ by the LLN}}$$

Generalization and uniform convergence

$$L(g^\star_{\mathrm{erm}}) - \inf_g L(g) = \underbrace{\left(L(g^\star_{\mathrm{erm}}) - \inf_{g\in\mathcal{G}} L(g)\right)}_{\text{Estimation error}} + \underbrace{\left(\inf_{g\in\mathcal{G}} L(g) - \inf_g L(g)\right)}_{\substack{\text{Approximation error}\\ \text{small if } \mathcal{G} \text{ rich enough}}}$$

$g^\star \in \operatorname*{Argmin}_{g\in\mathcal{G}} L(g) \neq \emptyset$

$$L(g^\star_{\mathrm{erm}}) - L(g^\star) = \overbrace{\left(L(g^\star_{\mathrm{erm}}) - \widehat{L}(g^\star_{\mathrm{erm}})\right)}^{\substack{\text{More complicated}\\ \text{Biased estimate}}} + \overbrace{\left(\widehat{L}(g^\star_{\mathrm{erm}}) - \widehat{L}(g^\star)\right)}^{\leq 0 \text{ by optimality}} + \overbrace{\left(\widehat{L}(g^\star) - L(g^\star)\right)}^{\to 0 \text{ by the LLN}}$$

Generalization and uniform convergence

$$L(g^\star_{\mathrm{erm}}) - \inf_g L(g) = \underbrace{\left(L(g^\star_{\mathrm{erm}}) - \inf_{g\in\mathcal{G}} L(g)\right)}_{\text{Estimation error}} + \underbrace{\left(\inf_{g\in\mathcal{G}} L(g) - \inf_g L(g)\right)}_{\substack{\text{Approximation error}\\ \text{small if } \mathcal{G} \text{ rich enough}}}$$

$g^\star \in \operatorname*{Argmin}_{g\in\mathcal{G}} L(g) \neq \emptyset$

$$L(g^\star_{\mathrm{erm}}) - L(g^\star) = \overbrace{\left(L(g^\star_{\mathrm{erm}}) - \widehat{L}(g^\star_{\mathrm{erm}})\right)}^{\substack{\text{More complicated}\\ \text{Biased estimate}}} + \overbrace{\left(\widehat{L}(g^\star_{\mathrm{erm}}) - \widehat{L}(g^\star)\right)}^{\leq 0 \text{ by optimality}} + \overbrace{\left(\widehat{L}(g^\star) - L(g^\star)\right)}^{\to 0 \text{ by the LLN}}$$

$$\leq \sup_{g\in\mathcal{G}} \left(L(g) - \widehat{L}(g)\right) + \sup_{g\in\mathcal{G}} \left(\widehat{L}(g) - L(g)\right)$$

Uniform bound

$$\leq 2 \sup_{g\in\mathcal{G}} |\widehat{L}(g) - L(g)|.$$

Generalization and uniform convergence

$$L(g^\star_{\mathrm{erm}}) - \inf_g L(g) = \underbrace{\left(L(g^\star_{\mathrm{erm}}) - \inf_{g\in\mathcal{G}} L(g)\right)}_{\text{Estimation error}} + \underbrace{\left(\inf_{g\in\mathcal{G}} L(g) - \inf_g L(g)\right)}_{\substack{\text{Approximation error}\\ \text{small if } \mathcal{G} \text{ rich enough}}}$$

$g^\star \in \operatorname*{Argmin}_{g\in\mathcal{G}} L(g) \neq \emptyset$

$$L(g^\star_{\mathrm{erm}}) - L(g^\star) = \overbrace{\left(L(g^\star_{\mathrm{erm}}) - \widehat{L}(g^\star_{\mathrm{erm}})\right)}^{\substack{\text{More complicated}\\ \text{Biased estimate}}} + \overbrace{\left(\widehat{L}(g^\star_{\mathrm{erm}}) - \widehat{L}(g^\star)\right)}^{\leq 0 \text{ by optimality}} + \overbrace{\left(\widehat{L}(g^\star) - L(g^\star)\right)}^{\to 0 \text{ by the LLN}}$$

$$\leq \sup_{g\in\mathcal{G}}\left(L(g) - \widehat{L}(g)\right) + \sup_{g\in\mathcal{G}}\left(\widehat{L}(g) - L(g)\right)$$

Uniform bound

$$\leq 2\sup_{g\in\mathcal{G}} |\widehat{L}(g) - L(g)|.$$

This bound intimately linked to **Rademacher complexity** of the **loss class** $\ell_{\mathcal{G}} \stackrel{\mathrm{def}}{=} \{(u,v) \mapsto \ell(v, g(u)) : \ g \in \mathcal{G}\}$

$$R_n(\ell_{\mathcal{G}}) \stackrel{\mathrm{def}}{=} \mathbb{E}\left[\sup_{\ell\in\ell_{\mathcal{G}}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \ell(v_i, g(u_i))\right]$$

ϵ_i i.i.d Rademacher (uniform on $\{\pm 1\}$).

Generalization and uniform convergence

$$L(g^\star_{\mathrm{erm}}) - \inf_g L(g) = \underbrace{\left(L(g^\star_{\mathrm{erm}}) - \inf_{g\in\mathcal{G}} L(g)\right)}_{\text{Estimation error}} + \underbrace{\left(\inf_{g\in\mathcal{G}} L(g) - \inf_g L(g)\right)}_{\substack{\text{Approximation error}\\ \text{small if } \mathcal{G} \text{ rich enough}}}$$

$g^\star \in \operatorname*{Argmin}_{g\in\mathcal{G}} L(g) \neq \emptyset$

$$\begin{aligned} L(g^\star_{\mathrm{erm}}) - L(g^\star) &= \overbrace{\left(L(g^\star_{\mathrm{erm}}) - \widehat{L}(g^\star_{\mathrm{erm}})\right)}^{\substack{\text{More complicated}\\ \text{Biased estimate}}} + \overbrace{\left(\widehat{L}(g^\star_{\mathrm{erm}}) - \widehat{L}(g^\star)\right)}^{\leq 0 \text{ by optimality}} + \overbrace{\left(\widehat{L}(g^\star) - L(g^\star)\right)}^{\to 0 \text{ by the LLN}} \\ &\leq \sup_{g\in\mathcal{G}}\left(L(g) - \widehat{L}(g)\right) + \sup_{g\in\mathcal{G}}\left(\widehat{L}(g) - L(g)\right) \\ &\leq 2\sup_{g\in\mathcal{G}} |\widehat{L}(g) - L(g)|. \end{aligned}$$

Uniform bound

This bound intimately linked to **Rademacher complexity** of the **loss class** $\ell_{\mathcal{G}} \stackrel{\mathrm{def}}{=} \{(u,v) \mapsto \ell(v, g(u)) : \ g \in \mathcal{G}\}$

$$R_n(\ell_{\mathcal{G}}) \stackrel{\mathrm{def}}{=} \mathbb{E}\left[\sup_{\ell\in\ell_{\mathcal{G}}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \ell(v_i, g(u_i))\right] \qquad \epsilon_i \text{ i.i.d Rademacher (uniform on } \{\pm 1\}).$$

- Expectation wrt data $(u_i, v_i)_{i\in[n]}$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$.
- Rademacher complexity measures the complexity of the function class $\ell_{\mathcal{G}}$ over n data points.
- It should converge to zero as n gets large, and this determines the generalization error rate.
- Can be estimated for some loss classes.
- Here $R_n(\ell_{\mathcal{G}})$ is the Rademacher complexity loss class $\ell_{\mathcal{G}}$, not $\mathcal{G}$. We can connect $R_n(\ell_{\mathcal{G}})$ to $R_n(\mathcal{G})$ in many cases : e.g. Lipschitz continuous loss, $0-1$ loss and binary classification.

Generalization and uniform convergence

Theorem *If ℓ is bounded on $\mathcal{G}$, then with probability at least $1-\delta$ on $(u_i, v_i)_{i\in[n]}$*

$$L(g^\star_{\mathrm{erm}}) - L(g^\star) \leq 4R_n(\ell_{\mathcal{G}}) + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Thus $L(g^\star_{\mathrm{erm}}) - L(g^\star) \underset{a.s.}{\to} 0$ if $R_n(\ell_{\mathcal{G}}) \underset{a.s.}{\to} 0$.

Generalization and uniform convergence

Theorem *If ℓ is bounded on $\mathcal{G}$, then with probability at least $1-\delta$ on $(u_i, v_i)_{i\in[n]}$*

$$L(g^\star_{\mathrm{erm}}) - L(g^\star) \leq 4R_n(\ell_{\mathcal{G}}) + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Thus $L(g^\star_{\mathrm{erm}}) - L(g^\star) \underset{a.s.}{\to} 0$ if $R_n(\ell_{\mathcal{G}}) \underset{a.s.}{\to} 0$.

- Typically a slow rate $O(1/\sqrt{n})$ (e.g. for Lipschitz losses $R_n(\ell_{\mathcal{G}}) = O(1/\sqrt{n})$).
- Can be improved to the faster rate $O(1/n)$ with a more refined analysis, under additional assumptions on the loss (e.g. strong convexity) and a suitably simple class $\mathcal{G}$.

Generalization and uniform convergence

Theorem *If ℓ is bounded on $\mathcal{G}$, then with probability at least $1-\delta$ on $(u_i, v_i)_{i\in[n]}$*

$$L(g^\star_{\mathrm{erm}}) - L(g^\star) \leq 4R_n(\ell_{\mathcal{G}}) + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Thus $L(g^\star_{\mathrm{erm}}) - L(g^\star) \underset{a.s.}{\to} 0$ if $R_n(\ell_{\mathcal{G}}) \underset{a.s.}{\to} 0$.

- Typically a slow rate $O(1/\sqrt{n})$ (e.g. for Lipschitz losses $R_n(\ell_{\mathcal{G}}) = O(1/\sqrt{n})$).
- Can be improved to the faster rate $O(1/n)$ with a more refined analysis, under additional assumptions on the loss (e.g. strong convexity) and a suitably simple class $\mathcal{G}$.
- Let $(g_k)_{k\in\mathbb{N}}$ be a sequence of iterates of an optimization algorithm to minimize L. Then,

$$\begin{aligned} L(g_k) - L(g^\star) &\leq \left(L(g_k) - \widehat{L}(g_k)\right) + \left(\widehat{L}(g_k) - \widehat{L}(g^\star_{\mathrm{erm}})\right) + \left(\widehat{L}(g^\star_{\mathrm{erm}}) - L(g^\star_{\mathrm{erm}})\right) + (L(g^\star_{\mathrm{erm}}) - L(g^\star)) \\ &\leq 4\sup_{g\in\mathcal{G}} |\widehat{L}(g) - L(g)| + \left(\widehat{L}(g_k) - \widehat{L}(g^\star_{\mathrm{erm}})\right) \end{aligned}$$

Generalization and uniform convergence

Theorem *If ℓ is bounded on $\mathcal{G}$, then with probability at least $1-\delta$ on $(u_i, v_i)_{i\in[n]}$*

$$L(g^\star_{\mathrm{erm}}) - L(g^\star) \leq 4R_n(\ell_{\mathcal{G}}) + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Thus $L(g^\star_{\mathrm{erm}}) - L(g^\star) \underset{a.s.}{\to} 0$ if $R_n(\ell_{\mathcal{G}}) \underset{a.s.}{\to} 0$.

- Typically a slow rate $O(1/\sqrt{n})$ (e.g. for Lipschitz losses $R_n(\ell_{\mathcal{G}}) = O(1/\sqrt{n})$).
- Can be improved to the faster rate $O(1/n)$ with a more refined analysis, under additional assumptions on the loss (e.g. strong convexity) and a suitably simple class $\mathcal{G}$.
- Let $(g_k)_{k\in\mathbb{N}}$ be a sequence of iterates of an optimization algorithm to minimize L. Then,

$$\begin{aligned} L(g_k) - L(g^\star) &\leq \left(L(g_k) - \widehat{L}(g_k)\right) + \left(\widehat{L}(g_k) - \widehat{L}(g^\star_{\mathrm{erm}})\right) + \left(\widehat{L}(g^\star_{\mathrm{erm}}) - L(g^\star_{\mathrm{erm}})\right) + (L(g^\star_{\mathrm{erm}}) - L(g^\star)) \\ &\leq 4\underbrace{\sup_{g\in\mathcal{G}} |\widehat{L}(g) - L(g)|}_{\text{Estimation error } O(1/\sqrt{n})} + \left(\widehat{L}(g_k) - \widehat{L}(g^\star_{\mathrm{erm}})\right) \end{aligned}$$

Generalization and uniform convergence

Theorem *If ℓ is bounded on $\mathcal{G}$, then with probability at least $1-\delta$ on $(u_i, v_i)_{i\in[n]}$*

$$L(g^\star_{\mathrm{erm}}) - L(g^\star) \leq 4R_n(\ell_{\mathcal{G}}) + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Thus $L(g^\star_{\mathrm{erm}}) - L(g^\star) \underset{a.s.}{\to} 0$ if $R_n(\ell_{\mathcal{G}}) \underset{a.s.}{\to} 0$.

- Typically a slow rate $O(1/\sqrt{n})$ (e.g. for Lipschitz losses $R_n(\ell_{\mathcal{G}}) = O(1/\sqrt{n})$).
- Can be improved to the faster rate $O(1/n)$ with a more refined analysis, under additional assumptions on the loss (e.g. strong convexity) and a suitably simple class $\mathcal{G}$.
- Let $(g_k)_{k\in\mathbb{N}}$ be a sequence of iterates of an optimization algorithm to minimize L. Then,

$$\begin{aligned} L(g_k) - L(g^\star) &\leq \left(L(g_k) - \widehat{L}(g_k)\right) + \left(\widehat{L}(g_k) - \widehat{L}(g^\star_{\mathrm{erm}})\right) + \left(\widehat{L}(g^\star_{\mathrm{erm}}) - L(g^\star_{\mathrm{erm}})\right) + \left(L(g^\star_{\mathrm{erm}}) - L(g^\star)\right) \\ &\leq \underbrace{4\sup_{g\in\mathcal{G}} |\widehat{L}(g) - L(g)|}_{\text{Estimation error } O(1/\sqrt{n})} + \underbrace{\left(\widehat{L}(g_k) - \widehat{L}(g^\star_{\mathrm{erm}})\right)}_{\text{Optimization error}} \end{aligned}$$

Generalization and uniform convergence

Theorem *If ℓ is bounded on $\mathcal{G}$, then with probability at least $1-\delta$ on $(u_i, v_i)_{i\in[n]}$*

$$L(g^\star_{\mathrm{erm}}) - L(g^\star) \leq 4R_n(\ell_{\mathcal{G}}) + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Thus $L(g^\star_{\mathrm{erm}}) - L(g^\star) \underset{a.s.}{\to} 0$ if $R_n(\ell_{\mathcal{G}}) \underset{a.s.}{\to} 0$.

- Typically a slow rate $O(1/\sqrt{n})$ (e.g. for Lipschitz losses $R_n(\ell_{\mathcal{G}}) = O(1/\sqrt{n})$).
- Can be improved to the faster rate $O(1/n)$ with a more refined analysis, under additional assumptions on the loss (e.g. strong convexity) and a suitably simple class $\mathcal{G}$.
- Let $(g_k)_{k\in\mathbb{N}}$ be a sequence of iterates of an optimization algorithm to minimize L. Then,

$$\begin{aligned} L(g_k) - L(g^\star) &\leq \left(L(g_k) - \widehat{L}(g_k)\right) + \left(\widehat{L}(g_k) - \widehat{L}(g^\star_{\mathrm{erm}})\right) + \left(\widehat{L}(g^\star_{\mathrm{erm}}) - L(g^\star_{\mathrm{erm}})\right) + (L(g^\star_{\mathrm{erm}}) - L(g^\star)) \\ &\leq \underbrace{4\sup_{g\in\mathcal{G}} |\widehat{L}(g) - L(g)|}_{\text{Estimation error } O(1/\sqrt{n})} + \underbrace{\left(\widehat{L}(g_k) - \widehat{L}(g^\star_{\mathrm{erm}})\right)}_{\text{Optimization error}} \end{aligned}$$

Take away messages

In machine learning, no need to optimize below estimation error,

i.e. up to accuracy $O(1/\sqrt{n})$ on the risk

Generalization and uniform convergence

Proof: We provide the proof of the upper-bound in expectation and the bound in probability follows from standard concentration arguments, for instance McDiarmid's inequality. The proof relies on a symmetrization trick. Let $D' = (u_i', v_i')_{i\in[n]}$ be an independent copy of the data $D = (u_i, v_i)_{i\in[n]}$, with corresponding empirical risk $\widehat{L}'(g) = \frac{1}{n}\sum_{i=1}^n \ell(v_i', g(u_i'))$. We have

$\mathbb{E}\left[\ell(v_i', g(u_i'))|D\right] = L(g)$

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} \widehat{L}(g) - L(g)\right] = \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\ell(v_i, g(u_i)) - \ell(v_i', g(u_i'))|D\right]\right]$$

$\sup \mathbb{E} \le \mathbb{E} \sup$

$$\le \mathbb{E}\left[\mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \left(\ell(v_i, g(u_i)) - \ell(v_i', g(u_i'))\right)|D\right]\right]$$

$\mathbb{E}\left[\mathbb{E}\left[\cdot|D\right]\right] = \mathbb{E}\left[\cdot\right]$

$$= \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \left(\ell(v_i, g(u_i)) - \ell(v_i', g(u_i'))\right)\right].$$

If X and X' are two i.i.d. r.v.'s, then $X \stackrel{d}{=} X'$ ($\stackrel{d}{=}$ is equality in distribution), and thus $h(X) \stackrel{d}{=} h(X')$ in distribution for any function h. Further, $h(X) - h(X') \stackrel{d}{=} h(X') - h(X) \stackrel{d}{=} -1\cdot(h(X) - h(X')) \stackrel{d}{=} \epsilon(h(X) - h(X'))$, where ϵ is uniform on $\{\pm 1\}$ and is independent of X and X'. Applying this to (u_i, v_i) and (u_i', v_i'), we get

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} \widehat{L}(g) - L(g)\right] \le \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \epsilon_i\left(\ell(v_i, g(u_i)) - \ell(v_i', g(u_i'))\right)\right]$$

$\sup$ is sublinear ϵ_i symmetric

$$\le \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \epsilon_i\left(\ell(v_i, g(u_i))\right)\right] + \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n -\epsilon_i \ell(v_i', g(u_i'))\right] = 2R_n(\ell_{\mathcal{G}}).$$

The same bound also applies to $L(g) - \widehat{L}(g)$, and recalling that $L(g^\star_{\text{erm}}) - L(g^\star) \le \sup_{g\in\mathcal{G}} \widehat{L}(g) - L(g) + \sup_{g\in\mathcal{G}} L(g) - \widehat{L}(g)$, we get

$$\mathbb{E}\left[L(g^\star_{\text{erm}}) - L(g^\star)\right] \le 4R_n(\ell_{\mathcal{G}}).$$

a.s. convergence follows by taking $\delta = 2/n^2$ and use Borel-Cantelli lemma.

Example: least-squares

$\ell(v, x^\top \varphi(u)) = \frac{1}{2}(v - x^\top \varphi(u))^2$, and thus

$$L(x) = \frac{1}{2}\mathbb{E}\left[(v - x^\top \varphi(u))^2\right] \quad \widehat{L}(x) = \frac{1}{2n}\sum_{i=1}^{n}(v_i - x^\top \varphi(u_i))^2.$$

We have

$$\widehat{L}(x) - L(x) = \frac{1}{2}x^\top \left(\frac{1}{n}\sum_{i=1}^{n}\varphi(u_i)\varphi(u_i)^\top - \mathbb{E}\left[\varphi(u)\varphi(u)^\top\right]\right) x$$
$$- x^\top \left(\frac{1}{n}\sum_{i=1}^{n} v_i\varphi(u_i) - \mathbb{E}\left[v\varphi(u)\right]\right) + \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} v_i^2 - \mathbb{E}\left[v^2\right]\right).$$

Thus

$$\sup_{\|x\|\leq c} |\widehat{L}(x) - L(x)| \leq \frac{c^2}{2}\left\|\left|\frac{1}{n}\sum_{i=1}^{n}\varphi(u_i)\varphi(u_i)^\top - \mathbb{E}\left[\varphi(u)\varphi(u)^\top\right]\right|\right\|$$
$$+ c\left\|\frac{1}{n}\sum_{i=1}^{n} v_i\varphi(u_i) - \mathbb{E}\left[v\varphi(u)\right]\right\| + \frac{1}{2}\left|\frac{1}{n}\sum_{i=1}^{n} v_i^2 - \mathbb{E}\left[v^2\right]\right|$$
$$= O(1/\sqrt{n})$$

with high probability from concentration of the empirical mean

Example: mean estimation

$$\widehat{L}(x) = \frac{1}{2n}\sum_{i=1}^{n}(x - v_i)^2 \qquad L(x) = \frac{1}{2}\mathbb{E}(x - v)^2$$

$$x^\star_{\text{erm}} = \operatorname*{Argmin}_{x\in\mathbb{R}} \widehat{L}(x) = \frac{1}{n}\sum_{i=1}^{n} v_i \qquad x^\star = \operatorname*{Argmin}_{x\in\mathbb{R}} L(x) = \mathbb{E}[v].$$

Example: mean estimation

$$\widehat{L}(x) = \frac{1}{2n}\sum_{i=1}^{n}(x - v_i)^2 \qquad L(x) = \frac{1}{2}\mathbb{E}(x - v)^2$$

$$x^\star_{\text{erm}} = \operatorname*{Argmin}_{x\in\mathbb{R}} \widehat{L}(x) = \frac{1}{n}\sum_{i=1}^{n} v_i \qquad x^\star = \operatorname*{Argmin}_{x\in\mathbb{R}} L(x) = \mathbb{E}[v].$$

$$\begin{aligned} L(x^\star_{\text{erm}}) - L(x^\star) &= \frac{(x^\star_{\text{erm}})^2}{2} - \frac{(x^\star)^2}{2} - \mathbb{E}[v](x^\star_{\text{erm}} - x^\star) \\ &= \frac{(x^\star_{\text{erm}})^2}{2} - \frac{E[v]^2}{2} - \mathbb{E}[v](x^\star_{\text{erm}} - \mathbb{E}[v]) \\ &= \frac{(x^\star_{\text{erm}})^2}{2} - \mathbb{E}[v]x^\star_{\text{erm}} + \frac{\mathbb{E}[v]^2}{2} = \frac{1}{2}(x^\star_{\text{erm}} - \mathbb{E}[v])^2 \\ &= \frac{1}{2}(x^\star_{\text{erm}} - \mathbb{E}[x^\star_{\text{erm}}])^2. \end{aligned}$$

iid samples

Example: mean estimation

$$\widehat{L}(x) = \frac{1}{2n}\sum_{i=1}^{n}(x - v_i)^2 \qquad L(x) = \frac{1}{2}\mathbb{E}(x - v)^2$$

$$x^\star_{\text{erm}} = \operatorname*{Argmin}_{x\in\mathbb{R}} \widehat{L}(x) = \frac{1}{n}\sum_{i=1}^{n} v_i \qquad x^\star = \operatorname*{Argmin}_{x\in\mathbb{R}} L(x) = \mathbb{E}[v].$$

$$\begin{aligned} L(x^\star_{\text{erm}}) - L(x^\star) &= \frac{(x^\star_{\text{erm}})^2}{2} - \frac{(x^\star)^2}{2} - \mathbb{E}[v](x^\star_{\text{erm}} - x^\star) \\ &= \frac{(x^\star_{\text{erm}})^2}{2} - \frac{E[v]^2}{2} - \mathbb{E}[v](x^\star_{\text{erm}} - \mathbb{E}[v]) \\ &= \frac{(x^\star_{\text{erm}})^2}{2} - \mathbb{E}[v]x^\star_{\text{erm}} + \frac{\mathbb{E}[v]^2}{2} = \frac{1}{2}(x^\star_{\text{erm}} - \mathbb{E}[v])^2 \\ &= \frac{1}{2}(x^\star_{\text{erm}} - \mathbb{E}[x^\star_{\text{erm}}])^2. \end{aligned}$$

iid samples

$$\mathbb{E}\left[L(x^\star_{\text{erm}}) - L(x^\star)\right] = \frac{1}{2}\text{Var}[x^\star_{\text{erm}}] = \frac{1}{2}\text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} v_i\right] = \frac{1}{2n^2}\sum_{i=1}^{n}\text{Var}[v_i] = \frac{\text{Var}[v]}{2n}.$$

iid samples

Example: mean estimation

$$\widehat{L}(x) = \frac{1}{2n}\sum_{i=1}^{n}(x - v_i)^2 \qquad L(x) = \frac{1}{2}\mathbb{E}(x - v)^2$$

$$x^\star_{\text{erm}} = \operatorname*{Argmin}_{x\in\mathbb{R}} \widehat{L}(x) = \frac{1}{n}\sum_{i=1}^{n} v_i \qquad x^\star = \operatorname*{Argmin}_{x\in\mathbb{R}} L(x) = \mathbb{E}[v].$$

$$\begin{aligned} L(x^\star_{\text{erm}}) - L(x^\star) &= \frac{(x^\star_{\text{erm}})^2}{2} - \frac{(x^\star)^2}{2} - \mathbb{E}[v](x^\star_{\text{erm}} - x^\star) \\ &= \frac{(x^\star_{\text{erm}})^2}{2} - \frac{E[v]^2}{2} - \mathbb{E}[v](x^\star_{\text{erm}} - \mathbb{E}[v]) \\ &= \frac{(x^\star_{\text{erm}})^2}{2} - \mathbb{E}[v]x^\star_{\text{erm}} + \frac{\mathbb{E}[v]^2}{2} = \frac{1}{2}(x^\star_{\text{erm}} - \mathbb{E}[v])^2 \\ &= \frac{1}{2}(x^\star_{\text{erm}} - \mathbb{E}[x^\star_{\text{erm}}])^2. \end{aligned}$$

iid samples

$$\mathbb{E}\left[L(x^\star_{\text{erm}}) - L(x^\star)\right] = \frac{1}{2}\text{Var}[x^\star_{\text{erm}}] = \frac{1}{2}\text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} v_i\right] = \frac{1}{2n^2}\sum_{i=1}^{n}\text{Var}[v_i] = \frac{\text{Var}[v]}{2n}.$$

iid samples

- Improved $O(1/n)$ rate.
- Valid only at $x^\star_{\text{erm}}$ (non-uniform) and uses strong convexity of the loss.

Machine learning for big data

$$\min_{x\in\mathbb{R}^d}\Big\{f(x)=\mathbb{E}_\xi[\ell(x,\xi)]\Big\}.$$

Machine learning for big data

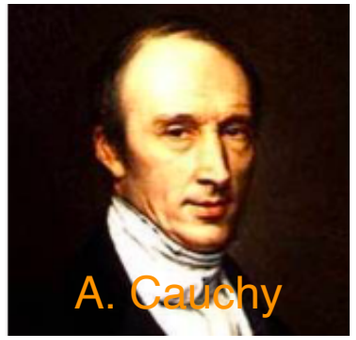
$$\min_{x\in\mathbb{R}^d}\left\{f(x)=\mathbb{E}_\xi[\ell(x,\xi)]\right\}.$$

- Large-scale machine learning: large d, large n :
  - d : dimension of each (input) observation.
  - n : number of observations.

Machine learning for big data

$$\min_{x\in\mathbb{R}^d}\Big\{f(x)=\mathbb{E}_\xi[\ell(x,\xi)]\Big\}.$$

- Large-scale machine learning: large d, large n :
  - d : dimension of each (input) observation.
  - n : number of observations.
- Gradient descent and variants: running-time complexity: $O(dn)$.


A. Cauchy


I. Gelfand

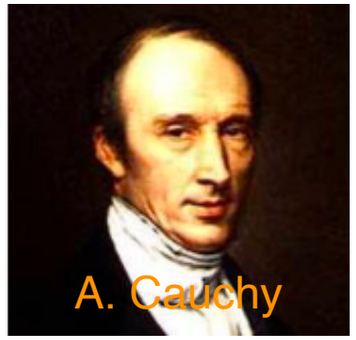

Y. Nesterov


H. Attouch

Machine learning for big data

$$\min_{x\in\mathbb{R}^d}\left\{f(x)=\mathbb{E}_\xi[\ell(x,\xi)]\right\}.$$

- Large-scale machine learning: large d, large n :
  - d : dimension of each (input) observation.
  - n : number of observations.
- Gradient descent and variants: running-time complexity: $O(dn)$.


A. Cauchy


I. Gelfand


Y. Nesterov

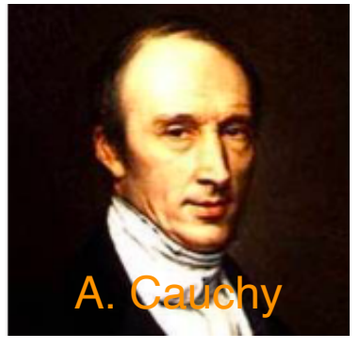

H. Attouch

- Scaling to large-scale problems: large d and large n.

Machine learning for big data

$$\min_{x\in\mathbb{R}^d}\Big\{f(x)=\mathbb{E}_\xi[\ell(x,\xi)]\Big\}.$$

- Large-scale machine learning: large d, large n :
  - d : dimension of each (input) observation.
  - n : number of observations.
- Gradient descent and variants: running-time complexity: $O(dn)$.

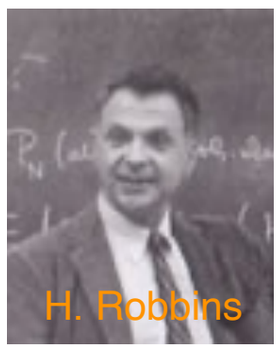

A. Cauchy

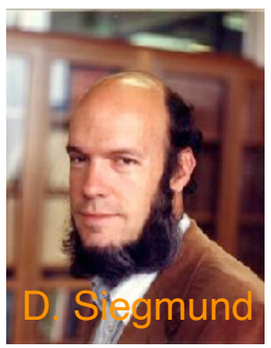

I. Gelfand


Y. Nesterov


H. Attouch

- Scaling to large-scale problems: large d and large n.
- Rather sample gradients: stochastic gradient descent and variants.


H. Robbins


D. Siegmund


S. Monro

Machine learning for big data

- Large-scale machine learning: large d, large n :
  - d : dimension of each (input) observation.
  - n : number of observations.
- Ideal running-time complexity: $O(dn)$.
- Scaling to large problems:
  - 1950's: computers not powerful enough.
  - 2010's: data too massive.
- Going back to stochastic methods for training:
  - Stochasticity on the data: stochastic gradient methods [Robbins and Monro, 1951].
  - Stochasticity on the decision variable: stochastic incremental methods, coordinate descent methods (not considered in this class).
  - Distributed methods: computation on distributed agents/units (not considered in this class)
  - Federated learning: data spread among several agents/clients who do not want to share/reveal them.

First-Order Stochastic Optimization

$$\min_{x \in \mathbb{R}^d} f(x).$$

- Population risk minimization :
  - Minimize $f(x) = \mathbb{E}_\xi \left[\ell(x, \xi)\right]$, $\xi \sim P$.

- Empirical risk minimization (special case of the above) :
  - Minimize $f(x) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(x)$.

First-Order Stochastic Optimization

$$\min_{x \in \mathbb{R}^d} f(x).$$

- Population risk minimization :
  - Minimize $f(x) = \mathbb{E}_\xi \left[\ell(x, \xi)\right]$, $\xi \sim P$.

- Empirical risk minimization (special case of the above) :
  - Minimize $f(x) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(x)$.

Input : algorithm parameters $(\eta_k)_{k \in \mathbb{N}}$, x_0, stopping rule, probability distributions $(P_k)_{k \in \mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic gradient estimate $G_k \sim P_k$;

  $x_{k+1} = h\left((x_i)_{i \leq k}, (G_i)_{i \leq k}, \eta_k\right)$;

  $k \leftarrow k + 1$.

return x_k.

First-Order Stochastic Optimization

$$\min_{x \in \mathbb{R}^d} f(x).$$

- Population risk minimization :
  - Minimize $f(x) = \mathbb{E}_\xi \left[\ell(x, \xi)\right]$, $\xi \sim P$.
  - Sample n iid samples $(\xi_i)_{i \in [n]}$ from P.
  - Take $G_k = \frac{1}{n} \sum_{i=1}^n \nabla \ell(x_k, \xi_i)$.
- Empirical risk minimization (special case of the above) :
  - Minimize $f(x) = \frac{1}{n} \sum_{i=1}^n \ell_i(x)$.
  - Sample a batch $B_k \subset [n]$.
  - Take $G_k = \frac{1}{|B_k|} \sum_{i \in B_k} \nabla \ell_i(x_k)$.

Input : algorithm parameters $(\eta_k)_{k \in \mathbb{N}}$, x_0, stopping rule, probability distributions $(P_k)_{k \in \mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic gradient estimate $G_k \sim P_k$;

  $x_{k+1} = h\left((x_i)_{i \leq k}, (G_i)_{i \leq k}, \eta_k\right)$;

  $k \leftarrow k + 1$.

return x_k.

First-Order Stochastic Algorithms

$$\min_{x \in \mathbb{R}^d} f(x).$$

Stochastic Gradient Descent (SGD)

Input : step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

Sample a stochastic gradient estimate $G_k \sim P_k$;

$x_{k+1} = x_k - \gamma_k G_k$;

$k \leftarrow k + 1$.

return x_k.

First-Order Stochastic Algorithms

$$\min_{x \in \mathbb{R}^d} f(x).$$

Stochastic Gradient Descent (SGD)

Input : step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic gradient estimate $G_k \sim P_k$;

  $x_{k+1} = x_k - \gamma_k G_k$;

  $k \leftarrow k + 1$.

return x_k.

Adaptive Gradient (AdaGrad)

Input : step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic gradient estimate $G_k \sim P_k$;

  $x_{k+1} = x_k - \gamma_k G_k / \sqrt{\sum_{i=0}^{k} G_k^2}$;

  $k \leftarrow k + 1$.

return x_k.

First-Order Stochastic Algorithms

$$\min_{x \in \mathbb{R}^d} f(x).$$

Stochastic Gradient Descent (SGD)

Input : step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic gradient estimate $G_k \sim P_k$;

  $x_{k+1} = x_k - \gamma_k G_k$;

  $k \leftarrow k + 1$.

return x_k.

Adaptive Gradient (AdaGrad)

Input : step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic gradient estimate $G_k \sim P_k$;

  $x_{k+1} = x_k - \gamma_k G_k / \sqrt{\sum_{i=0}^{k} G_k^2}$;

  $k \leftarrow k + 1$.

return x_k.

ADAptive Moment estimation (ADAM)

Input : $\gamma > 0$, $\varepsilon > 0$, $(\alpha, \beta) \in [0, 1[$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic gradient estimate $G_k \sim P_k$;

  $y_k = \alpha y_{k-1} + (1-\alpha) G_k$; (biased 1st moment estimate)

  $z_k = \beta z_{k-1} + (1-\beta) G_k^2$; (biased 2nd moment estimate)

  $\hat{y}_k = y_k / (1 - \alpha^k)$; bias correction of 1st moment

  $\hat{z}_k = z_k / (1 - \beta^k)$; bias correction of 2nd moment

  $x_{k+1} = x_k - \gamma \hat{y}_k / (\varepsilon + \sqrt{\hat{z}_k})$;

  $k \leftarrow k + 1$.

return x_k.

First-Order Stochastic Algorithms

$$\min_{x \in \mathbb{R}^d} f(x).$$

Stochastic Gradient Descent (SGD)

Input : step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic gradient estimate $G_k \sim P_k$;

  $x_{k+1} = x_k - \gamma_k G_k$;

  $k \leftarrow k + 1$.

return x_k.

Adaptive Gradient (AdaGrad)

Input : step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic gradient estimate $G_k \sim P_k$;

  $x_{k+1} = x_k - \gamma_k G_k / \sqrt{\sum_{i=0}^{k} G_k^2}$;

  $k \leftarrow k + 1$.

return x_k.

ADAptive Moment estimation (ADAM)

Input : $\gamma > 0$, $\varepsilon > 0$, $(\alpha, \beta) \in [0, 1[$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic gradient estimate $G_k \sim P_k$;

  $y_k = \alpha y_{k-1} + (1-\alpha) G_k$; (biased 1st moment estimate)

  $z_k = \beta z_{k-1} + (1-\beta) G_k^2$; (biased 2nd moment estimate)

  $\hat{y}_k = y_k / (1 - \alpha^k)$; bias correction of 1st moment

  $\hat{z}_k = z_k / (1 - \beta^k)$; bias correction of 2nd moment

  $x_{k+1} = x_k - \gamma \hat{y}_k / (\varepsilon + \sqrt{\hat{z}_k})$;

  $k \leftarrow k + 1$.

return x_k.

- These are the most popular.

First-Order Stochastic Algorithms

$$\min_{x \in \mathbb{R}^d} f(x).$$

Stochastic Gradient Descent (SGD)

Input : step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic gradient estimate $G_k \sim P_k$;

  $x_{k+1} = x_k - \gamma_k G_k$;

  $k \leftarrow k + 1$.

return x_k.

- These are the most popular.
- We will focus this class on SGD:
  - by far the most popular,
  - simpler to analyze,
  - give the key tools to understand and analyze other algorithms.

First-Order Stochastic Algorithms

$$\min_{x \in \mathbb{R}^d} f(x).$$

Stochastic Gradient Descent (SGD)

Input : step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic gradient estimate $G_k \sim P_k$;

  $x_{k+1} = x_k - \gamma_k G_k$;

  $k \leftarrow k + 1$.

return x_k.

- $G_k = \nabla f(x_k) + e_k$: e_k is the stochastic error on the gradient.
- The distribution of the error e_k must only depend on information up to iteration k.
- Given that information, e_k should have :
  - zero mean : i.e. G_k is an **unbiased** estimate of $\nabla f(x_k)$;
  - a controlled mean "amplitude" : i.e. the **variance** of G_k should be bounded in a certain way.
- If one hopes to ensure any convergence guarantee of the algorithm, either :
  - G_k has to eventually point to the right gradient : the variance should vanish, i.e. $G_k - \nabla f(x_k) \to 0$ in some stochastic sense ;
  - or use the step-sizes to cancel out the error term.
- The aim in the rest of the course is to give this firm theoretical grounds.

Outline

- Classes of functions.
- Toolbox on sequences.
- Deterministic smooth optimization.
- Stochastic approximation à la Robbins-Monro.
- Stochastic gradient descent: vanishing step-size.
- Stochastic gradient descent for finite sums.

Outline

- Classes of functions.
- Toolbox on sequences.
- Deterministic smooth optimization.
- Stochastic approximation à la Robbins-Monro.
- Stochastic gradient descent: vanishing step-size.
- Stochastic gradient descent for finite sums.

Differentiability

In the following, $\|\cdot\|$ is the euclidian norm on $\mathbb{R}^n$ for any n and the dimension is to be understood from the context.

Definition *We denote by $\mathcal{L}(\mathbb{R}^d, \mathbb{R}^m)$ the vector space of continuous linear operators from $\mathbb{R}^d$ to $\mathbb{R}^m$. It is endowed with the norm*

$$\left|\!\left|\!\left| A \right|\!\right|\!\right| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

Differentiability

In the following, $\|\cdot\|$ is the euclidian norm on $\mathbb{R}^n$ for any n and the dimension is to be understood from the context.

Definition *We denote by $\mathcal{L}(\mathbb{R}^d, \mathbb{R}^m)$ the vector space of continuous linear operators from $\mathbb{R}^d$ to $\mathbb{R}^m$. It is endowed with the norm*

$$\left\|\left\| A \right\|\right\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

Definition *A function $F : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}^m$, where Ω is an open subset, is (Fréchet) differentiable at $x \in \Omega$ if there exists an operator $F'(x) \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^m)$, called the (Fréchet) derivative of F at x, such that*

$$F(x+z) = F(x) + F'(x)z + o(\|z\|), \quad \forall z \in \mathbb{R}^d.$$

This element is unique, when it exists.

Differentiability

In the following, $\|\cdot\|$ is the euclidian norm on $\mathbb{R}^n$ for any n and the dimension is to be understood from the context.

Definition *We denote by $\mathcal{L}(\mathbb{R}^d, \mathbb{R}^m)$ the vector space of continuous linear operators from $\mathbb{R}^d$ to $\mathbb{R}^m$. It is endowed with the norm*

$$\left\|\left\|A\right\|\right\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

Definition *A function $F : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}^m$, where Ω is an open subset, is (Fréchet) differentiable at $x \in \Omega$ if there exists an operator $F'(x) \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^m)$, called the (Fréchet) derivative of F at x, such that*

$$F(x+z) = F(x) + F'(x)z + o(\|z\|), \quad \forall z \in \mathbb{R}^d.$$

This element is unique, when it exists.

Remark *Ω is supposed open to ensure uniqueness of the derivative.*

Differentiability

In the following, $\|\cdot\|$ is the euclidian norm on $\mathbb{R}^n$ for any n and the dimension is to be understood from the context.

Definition *We denote by $\mathcal{L}(\mathbb{R}^d, \mathbb{R}^m)$ the vector space of continuous linear operators from $\mathbb{R}^d$ to $\mathbb{R}^m$. It is endowed with the norm*

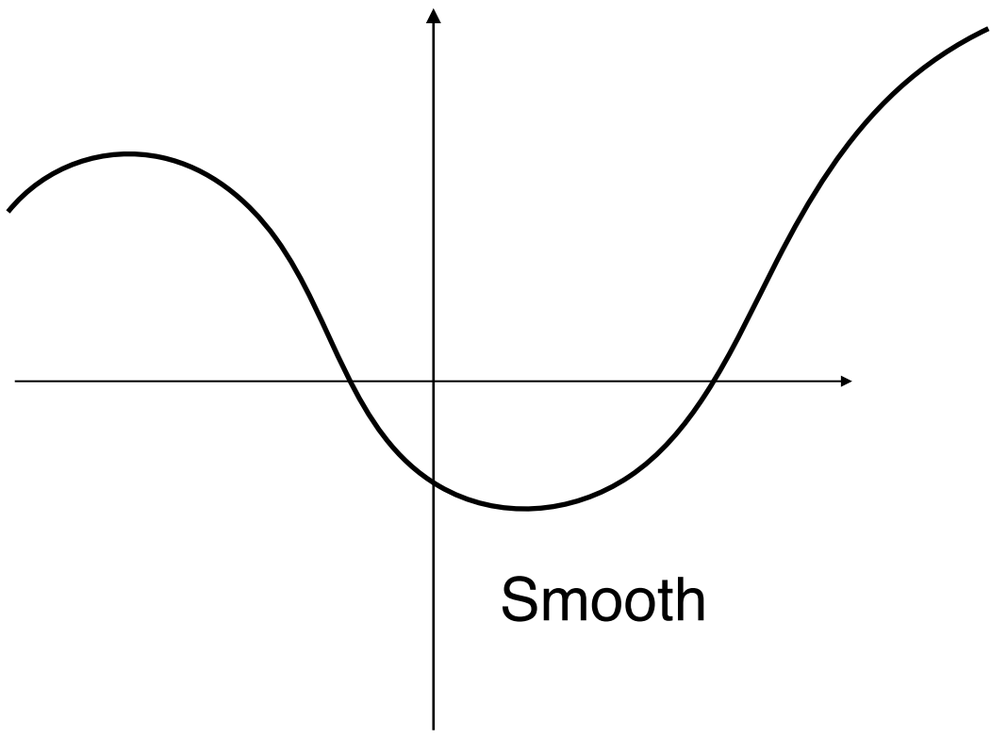
$$\left\|\left\| A \right\|\right\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

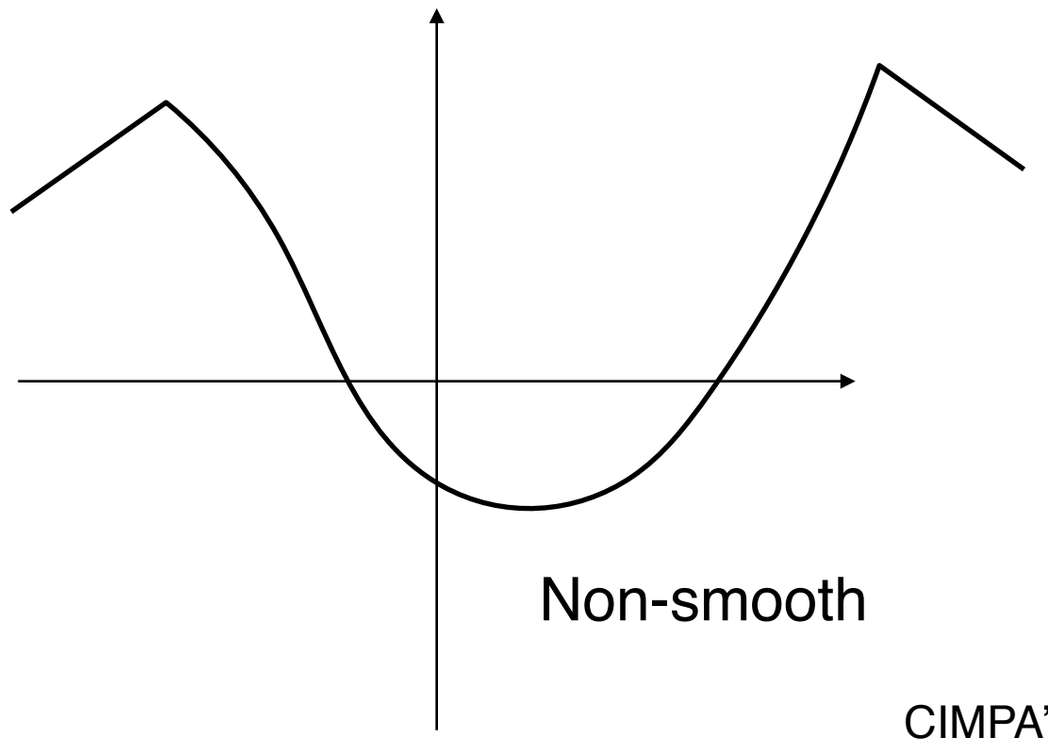
Definition *A function $F : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}^m$, where Ω is an open subset, is (Fréchet) differentiable at $x \in \Omega$ if there exists an operator $F'(x) \in \mathcal{L}(\mathbb{R}^d, \mathbb{R}^m)$, called the (Fréchet) derivative of F at x, such that*

$$F(x+z) = F(x) + F'(x)z + o(\|z\|), \quad \forall z \in \mathbb{R}^d.$$

This element is unique, when it exists.

Remark *Ω is supposed open to ensure uniqueness of the derivative.*





Differentiability

Definition *Let $F : \mathbb{R}^d \to \mathbb{R}^m$. If F is differentiable with respect to the k^{th} component of x (with the other components fixed), we denote $\frac{\partial F}{\partial x_k}(x)$ this derivative, which is called the partial derivative of F wrt the k^{th} variable*

Differentiability

Definition *Let $F : \mathbb{R}^d \to \mathbb{R}^m$. If F is differentiable with respect to the k^{th} component of x (with the other components fixed), we denote $\frac{\partial F}{\partial x_k}(x)$ this derivative, which is called the partial derivative of F wrt the k^{th} variable*

Proposition *If F is differentiable at x, then it has partial derivatives and*

$$F'(x)z = \sum_{k=1}^{d} \frac{\partial F}{\partial x_k}(x) z_k, \quad \forall z \in \mathbb{R}^d.$$

Differentiability

Definition *Let $F : \mathbb{R}^d \to \mathbb{R}^m$. If F is differentiable with respect to the k^{th} component of x (with the other components fixed), we denote $\frac{\partial F}{\partial x_k}(x)$ this derivative, which is called the partial derivative of F wrt the k^{th} variable*

Proposition *If F is differentiable at x, then it has partial derivatives and*

$$F'(x)z = \sum_{k=1}^{d} \frac{\partial F}{\partial x_k}(x) z_k, \quad \forall z \in \mathbb{R}^d.$$

Remark *The converse is not true. Take for example the function $F : (x_1, x_2) \in \mathbb{R}^2 \mapsto \begin{cases} 0 & \text{if } x_1 x_2 = 0 \\ 1 & \text{otherwise} \end{cases}$. The partial derivatives are both 0 at the origin, but the function is NOT differentiable at the origin since it is not even continuous there.*

Differentiability

Definition *Let $F : \mathbb{R}^d \to \mathbb{R}^m$. If F is differentiable with respect to the k^{th} component of x (with the other components fixed), we denote $\frac{\partial F}{\partial x_k}(x)$ this derivative, which is called the partial derivative of F wrt the k^{th} variable*

Proposition *If F is differentiable at x, then it has partial derivatives and*

$$F'(x)z = \sum_{k=1}^{d} \frac{\partial F}{\partial x_k}(x) z_k, \quad \forall z \in \mathbb{R}^d.$$

Remark *The converse is not true. Take for example the function $F : (x_1, x_2) \in \mathbb{R}^2 \mapsto \begin{cases} 0 & \text{if } x_1 x_2 = 0 \\ 1 & \text{otherwise} \end{cases}$. The partial derivatives are both 0 at the origin, but the function is NOT differentiable at the origin since it is not even continuous there.*

The derivative in this case is nothing but the Jacobian matrix

$$J_F(x) \stackrel{\text{def}}{=} F'(x) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(x) & \frac{\partial F_1}{\partial x_2}(x) & \cdots & \frac{\partial F_1}{\partial x_d}(x) \\ \frac{\partial F_2}{\partial x_1}(x) & \frac{\partial F_2}{\partial x_2}(x) & \cdots & \frac{\partial F_2}{\partial x_d}(x) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial F_m}{\partial x_1}(x) & \frac{\partial F_m}{\partial x_2}(x) & \cdots & \frac{\partial F_m}{\partial x_d}(x) \end{pmatrix}$$

Differentiability

Definition *Higher-order Fréchet derivatives are defined inductively. Thus, the second Fréchet derivative of* F *at* x *is the (bilinear) operator* $F''(x) \in \mathcal{L}(\mathbb{R}^d, \mathcal{L}(\mathbb{R}^d, \mathbb{R}^m))$ *which satisfies*

$$F'(x+z) = F'(x) + F''(x)z + o(\|z\|), \forall z \in \mathbb{R}^d.$$

Differentiability

Definition *Higher-order Fréchet derivatives are defined inductively. Thus, the second Fréchet derivative of* F *at* x *is the (bilinear) operator* $F''(x) \in \mathcal{L}(\mathbb{R}^d, \mathcal{L}(\mathbb{R}^d, \mathbb{R}^m))$ *which satisfies*

$$F'(x+z) = F'(x) + F''(x)z + o(\|z\|), \forall z \in \mathbb{R}^d.$$

Proposition *If* F'' *is twice differentiable at* x*, then* $F''(x)$ *is a symmetric bilinear operator.*

Proof: Bilinearity comes from the isomorphism between $\mathcal{L}(\mathbb{R}^d, \mathcal{L}(\mathbb{R}^d, \mathbb{R}^m))$ and the vector space of bilinear operators. Symmetry is a consequence of the fact that the roles of two directions when taking the derivative of the derivative are exchangeable. ■

Differentiability

Definition *Higher-order Fréchet derivatives are defined inductively. Thus, the second Fréchet derivative of F at x is the (bilinear) operator $F''(x) \in \mathcal{L}(\mathbb{R}^d, \mathcal{L}(\mathbb{R}^d, \mathbb{R}^m))$ which satisfies*

$$F'(x+z) = F'(x) + F''(x)z + o(\|z\|), \forall z \in \mathbb{R}^d.$$

Proposition *If F'' is twice differentiable at x, then $F''(x)$ is a symmetric bilinear operator.*

Proof: Bilinearity comes from the isomorphism between $\mathcal{L}(\mathbb{R}^d, \mathcal{L}(\mathbb{R}^d, \mathbb{R}^m))$ and the vector space of bilinear operators. Symmetry is a consequence of the fact that the roles of two directions when taking the derivative of the derivative are exchangeable. ■

Example *When $m = 1$, then*

$$F''(x)(z, y) = \sum_{i,j=1}^{d} \frac{\partial^2 F}{\partial x_i \partial x_j}(x) z_i y_j,$$

where $\frac{\partial^2 F}{\partial x_i \partial x_j}(x)$ are the second partial derivatives of F at x.

Chain rule

Theorem *Let $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}^m$ and $G : \Xi \subset \mathbb{R}^m \to \mathbb{R}^l$, where Ω and Ξ are open sets such that $F(\Omega) \subset \Xi$. Suppose that F is differentiable at $x \in \Omega$ and G is differentiable at $F(z) \in \Xi$. Then $G \circ F$ is differentiable at x with*

$$(G \circ F)'(x) = G'(F(x))F'(x).$$

Chain rule

Theorem *Let $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}^m$ and $G : \Xi \subset \mathbb{R}^m \to \mathbb{R}^l$, where Ω and Ξ are open sets such that $F(\Omega) \subset \Xi$. Suppose that F is differentiable at $x \in \Omega$ and G is differentiable at $F(z) \in \Xi$. Then $G \circ F$ is differentiable at x with*

$$(G \circ F)'(x) = G'(F(x))F'(x).$$

Remark *In terms of Jacobian matrices, the chain rule reads*

$$J_{G\circ F}(x) \stackrel{\text{def}}{=} \begin{pmatrix} \frac{\partial (G\circ F)_1}{\partial x_1}(x) & \frac{\partial (G\circ F)_1}{\partial x_2}(x) & \cdots & \frac{\partial (G\circ F)_1}{\partial x_d}(x) \\ \frac{\partial (G\circ F)_2}{\partial x_1}(x) & \frac{\partial (G\circ F)_2}{\partial x_2}(x) & \cdots & \frac{\partial (G\circ F)_2}{\partial x_d}(x) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial (G\circ F)_l}{\partial x_1}(x) & \frac{\partial (G\circ F)_l}{\partial x_2}(x) & \cdots & \frac{\partial (G\circ F)_l}{\partial x_d}(x) \end{pmatrix}$$

$$= J_G(F(x))J_F(x) \stackrel{\text{def}}{=} \begin{pmatrix} \frac{\partial G_1}{\partial z_1}(F(x)) & \frac{\partial G_1}{\partial z_2}F(x) & \cdots & \frac{\partial G_1}{\partial z_m}(F(x)) \\ \frac{\partial G_2}{\partial z_1}(F(x)) & \frac{\partial G_2}{\partial z_2}(F(x)) & \cdots & \frac{\partial G_2}{\partial z_m}(F(x)) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial G_l}{\partial z_1}(F(x)) & \frac{\partial G_l}{\partial z_2}(F(x)) & \cdots & \frac{\partial G_l}{\partial z_m}(F(x)) \end{pmatrix} \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(x) & \frac{\partial F_1}{\partial x_2}(x) & \cdots & \frac{\partial F_1}{\partial x_d}(x) \\ \frac{\partial F_2}{\partial x_1}(x) & \frac{\partial F_2}{\partial x_2}(x) & \cdots & \frac{\partial F_2}{\partial x_d}(x) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial F_m}{\partial x_1}(x) & \frac{\partial F_m}{\partial x_2}(x) & \cdots & \frac{\partial F_m}{\partial x_d}(x) \end{pmatrix}$$

Chain rule

Theorem *Let $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}^m$ and $G : \Xi \subset \mathbb{R}^m \to \mathbb{R}^l$, where Ω and Ξ are open sets such that $F(\Omega) \subset \Xi$. Suppose that F is differentiable at $x \in \Omega$ and G is differentiable at $F(z) \in \Xi$. Then $G \circ F$ is differentiable at x with*

$$(G \circ F)'(x) = G'(F(x))F'(x).$$

Remark *In terms of Jacobian matrices, the chain rule reads*

$$J_{G\circ F}(x) \stackrel{\text{def}}{=} \begin{pmatrix} \frac{\partial (G\circ F)_1}{\partial x_1}(x) & \frac{\partial (G\circ F)_1}{\partial x_2}(x) & \cdots & \frac{\partial (G\circ F)_1}{\partial x_d}(x) \\ \frac{\partial (G\circ F)_2}{\partial x_1}(x) & \frac{\partial (G\circ F)_2}{\partial x_2}(x) & \cdots & \frac{\partial (G\circ F)_2}{\partial x_d}(x) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial (G\circ F)_l}{\partial x_1}(x) & \frac{\partial (G\circ F)_l}{\partial x_2}(x) & \cdots & \frac{\partial (G\circ F)_l}{\partial x_d}(x) \end{pmatrix}$$

$$= J_G(F(x))J_F(x) \stackrel{\text{def}}{=} \begin{pmatrix} \frac{\partial G_1}{\partial z_1}(F(x)) & \frac{\partial G_1}{\partial z_2}F(x) & \cdots & \frac{\partial G_1}{\partial z_m}(F(x)) \\ \frac{\partial G_2}{\partial z_1}(F(x)) & \frac{\partial G_2}{\partial z_2}(F(x)) & \cdots & \frac{\partial G_2}{\partial z_m}(F(x)) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial G_l}{\partial z_1}(F(x)) & \frac{\partial G_l}{\partial z_2}(F(x)) & \cdots & \frac{\partial G_l}{\partial z_m}(F(x)) \end{pmatrix} \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(x) & \frac{\partial F_1}{\partial x_2}(x) & \cdots & \frac{\partial F_1}{\partial x_d}(x) \\ \frac{\partial F_2}{\partial x_1}(x) & \frac{\partial F_2}{\partial x_2}(x) & \cdots & \frac{\partial F_2}{\partial x_d}(x) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial F_m}{\partial x_1}(x) & \frac{\partial F_m}{\partial x_2}(x) & \cdots & \frac{\partial F_m}{\partial x_d}(x) \end{pmatrix}$$

Applies only to the smooth case,
though in ML/DL many apply it with non-smooth functions

Chain rule

Proof: We have

$$\begin{aligned} G(F(x+z)) &= G\Big(F(x)+F'(x)z+o(\|z\|\Big) \\ &= G(F(x))+G'(F(x))\Big(F'(x)z+o(\|z\|)\Big)+o\left(\|F'(x)z+o(\|z\|)\|\right) \\ &= G(F(x))+G'(F(x))F'(x)z+o\Big(G'(F(x))\,\|z\|\Big)+o\left(\|F'(x)z+o(\|z\|)\|\right). \end{aligned}$$

Since $F'(x)$ and $G'(F(x))$ are bounded linear operators, we have

$$o\Big(G'(F(x))\,\|z\|\Big)=o(\|z\|) \quad \text{and} \quad o\left(\|F'(x)z+o(\|z\|)\|\right)=o(\|z\|).$$

■

Gradient and Hessian

In optimization, we are primarily interested in functions $F : \mathbb{R}^d \to \mathbb{R}$. We equip $\mathbb{R}^n$ with an scalar product that we denote $\langle \cdot, \cdot \rangle$.

Definition (Gradient) *Let $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}$, and suppose that F is differentiable at $x \in \mathbb{R}^d$. The gradient of F at x is the vector denoted $\nabla F(x) \in \mathbb{R}^d$ such as*

$$F'(x)z = \langle \nabla F(x), z \rangle, \quad \forall z \in \mathbb{R}^d.$$

The gradient exists and is unique by the Riesz–Fréchet representation theorem.

Gradient and Hessian

In optimization, we are primarily interested in functions $F : \mathbb{R}^d \to \mathbb{R}$. We equip $\mathbb{R}^n$ with an scalar product that we denote $\langle \cdot, \cdot \rangle$.

Definition (Gradient) *Let $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}$, and suppose that F is differentiable at $x \in \mathbb{R}^d$. The gradient of F at x is the vector denoted $\nabla F(x) \in \mathbb{R}^d$ such as*

$$F'(x)z = \langle \nabla F(x), z \rangle, \quad \forall z \in \mathbb{R}^d.$$

The gradient exists and is unique by the Riesz–Fréchet representation theorem.

Definition (Hessian) *Let $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}$, and suppose that F is twice differentiable at $x \in \mathbb{R}^d$. The Hessian of F at x is the matrix denoted $\nabla^2 F(x) \in \mathbb{R}^{d \times d}$ such that*

$$F''(x)(z, y) = \left\langle z, \nabla^2 F(x) y \right\rangle = \left\langle y, \nabla^2 F(x) z \right\rangle, \quad \forall z, y \in \mathbb{R}^d.$$

Again, the Hessian exists and is unique and it is a symmetric matrix as $F''(x)$ is a symmetric operator.

Gradient and Hessian

In optimization, we are primarily interested in functions $F : \mathbb{R}^d \to \mathbb{R}$. We equip $\mathbb{R}^n$ with an scalar product that we denote $\langle \cdot, \cdot \rangle$.

Definition (Gradient) *Let $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}$, and suppose that F is differentiable at $x \in \mathbb{R}^d$. The gradient of F at x is the vector denoted $\nabla F(x) \in \mathbb{R}^d$ such as*

$$F'(x)z = \langle \nabla F(x), z \rangle, \quad \forall z \in \mathbb{R}^d.$$

The gradient exists and is unique by the Riesz–Fréchet representation theorem.

Definition (Hessian) *Let $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}$, and suppose that F is twice differentiable at $x \in \mathbb{R}^d$. The Hessian of F at x is the matrix denoted $\nabla^2 F(x) \in \mathbb{R}^{d \times d}$ such that*

$$F''(x)(z, y) = \left\langle z, \nabla^2 F(x) y \right\rangle = \left\langle y, \nabla^2 F(x) z \right\rangle, \quad \forall z, y \in \mathbb{R}^d.$$

Again, the Hessian exists and is unique and it is a symmetric matrix as $F''(x)$ is a symmetric operator.

Remark *The Hessian is the derivative of the gradient.*

Gradient and Hessian

- If we choose the scalar product $\langle x, z\rangle = x^\top z = z^\top x$, then from S26 and S27, we have

$$F'(x)z = \sum_{i=1}^{d} \frac{\partial F}{\partial x_i}(x) z_i = \nabla F(x)^\top z = \sum_{i=1}^{d} (\nabla F(x))_i z_i \quad \text{and}$$

$$F''(x)(z,y) = \sum_{i,j=1}^{d} \frac{\partial^2 F}{\partial x_i \partial x_j}(x) z_i y_j = z^\top \nabla^2 F(x) y = y^\top \nabla^2 F(x) z = \sum_{i,j=1}^{d} (\nabla^2 F(x))_{ij} z_i y_j,$$

which entails that

$$\nabla F(x) = \begin{pmatrix} \frac{\partial F}{\partial x_1}(x) \\ \frac{\partial F}{\partial x_2}(x) \\ \vdots \\ \frac{\partial F}{\partial x_d}(x) \end{pmatrix} \quad \text{and} \quad \nabla^2 F(x) = \begin{pmatrix} \frac{\partial^2 F}{\partial^2 x_1}(x) & \frac{\partial^2 F}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 F}{\partial x_1 \partial x_d}(x) \\ \frac{\partial^2 F}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 F}{\partial^2 x_2}(x) & \cdots & \frac{\partial^2 F}{\partial x_2 \partial x_d}(x) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 F}{\partial x_d \partial x_1}(x) & \frac{\partial^2 F}{\partial x_d \partial x_2}(x) & \cdots & \frac{\partial^2 F}{\partial^2 x_d}(x) \end{pmatrix}.$$

- The form of the gradient and the Hessian **depend on the choice of scalar product**, and the above choice is not the only one.
- In this course, we will essentially work with the scalar product $\langle x, z\rangle = x^\top z = z^\top x$.

Chain rule of gradient and Hessian

Theorem *Let $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}^m$ and $G : \Xi \subset \mathbb{R}^m \to \mathbb{R}$, where Ω and Ξ are open sets such that $F(\Omega) \subset \Xi$. Suppose that F is differentiable at $x \in \Omega$ and G is differentiable at $F(x) \in \Xi$. Then $G \circ F$ is differentiable at x with*

$$\nabla(G \circ F)(x) = J_F(x)^\top \nabla G(F(x)).$$

Chain rule of gradient and Hessian

Theorem *Let $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}^m$ and $G : \Xi \subset \mathbb{R}^m \to \mathbb{R}$, where Ω and Ξ are open sets such that $F(\Omega) \subset \Xi$. Suppose that F is differentiable at $x \in \Omega$ and G is differentiable at $F(x) \in \Xi$. Then $G \circ F$ is differentiable at x with*

$$\nabla(G \circ F)(x) = J_F(x)^\top \nabla G(F(x)).$$

Example *F is affine, i.e. $F(x) = Ax + b$, $A \in \mathbb{R}^{m \times d}, b \in \mathbb{R}^m$. Therefore,*

$$\nabla(G \circ F)(x) = A^\top \nabla G(Ax + b).$$

Chain rule of gradient and Hessian

Theorem *Let $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}^m$ and $G : \Xi \subset \mathbb{R}^m \to \mathbb{R}$, where Ω and Ξ are open sets such that $F(\Omega) \subset \Xi$. Suppose that F is differentiable at $x \in \Omega$ and G is differentiable at $F(x) \in \Xi$. Then $G \circ F$ is differentiable at x with*

$$\nabla(G \circ F)(x) = J_F(x)^\top \nabla G(F(x)).$$

Example *F is affine, i.e. $F(x) = Ax + b$, $A \in \mathbb{R}^{m \times d}, b \in \mathbb{R}^m$. Therefore,*

$$\nabla(G \circ F)(x) = A^\top \nabla G(Ax + b).$$

Theorem *Let $F : \mathbb{R}^d \to \mathbb{R}^m$ and $G : \mathbb{R}^m \to \mathbb{R}$ be twice differentiable at $x \in \mathbb{R}^d$ and $F(x) \in \mathbb{R}^m$ such that $G \circ F$ is twice differentiable at x. Then $G \circ F$ is twice differentiable at x with*

$$\nabla^2(G \circ F)(x) = J_F(x)^\top \nabla^2 G(F(x)) J_F(x) + \sum_{i=1}^{m} \frac{\partial G}{\partial z_i}(F(x)) \nabla^2 F_i(x).$$

Chain rule of gradient and Hessian

Theorem *Let $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}^m$ and $G : \Xi \subset \mathbb{R}^m \to \mathbb{R}$, where Ω and Ξ are open sets such that $F(\Omega) \subset \Xi$. Suppose that F is differentiable at $x \in \Omega$ and G is differentiable at $F(x) \in \Xi$. Then $G \circ F$ is differentiable at x with*

$$\nabla(G \circ F)(x) = J_F(x)^\top \nabla G(F(x)).$$

Example *F is affine, i.e. $F(x) = Ax + b$, $A \in \mathbb{R}^{m \times d}, b \in \mathbb{R}^m$. Therefore,*

$$\nabla(G \circ F)(x) = A^\top \nabla G(Ax + b).$$

Theorem *Let $F : \mathbb{R}^d \to \mathbb{R}^m$ and $G : \mathbb{R}^m \to \mathbb{R}$ be twice differentiable at $x \in \mathbb{R}^d$ and $F(x) \in \mathbb{R}^m$ such that $G \circ F$ is twice differentiable at x. Then $G \circ F$ is twice differentiable at x with*

$$\nabla^2(G \circ F)(x) = J_F(x)^\top \nabla^2 G(F(x)) J_F(x) + \sum_{i=1}^{m} \frac{\partial G}{\partial z_i}(F(x)) \nabla^2 F_i(x).$$

Example *F is affine, i.e. $F(x) = Ax + b$, $\mathbb{R}^{m \times d}, b \in \mathbb{R}^m$. Therefore,*

$$\nabla^2(G \circ F)(x) = A^\top \nabla^2 G(Ax + b) A.$$

Chain rule of gradient and Hessian

Proof: For the gradient, applying the chain rule of S28, we get

$$J_{G\circ F}(x) = J_G(F(x))J_F(x).$$

But by definition of the gradient, we have $J_{G\circ F}(x) = \nabla(G \circ F)(x)^\top$ and $J_G(F(x)) = \nabla G(F(x))^\top$. Taking the transpose, we conclude.

Let us turn to the Hessian, and recall that it the derivative of the gradient. Let $H(x) \stackrel{\text{def}}{=} \nabla(G \circ F)(x)$, we have

$$\begin{aligned}
H(x+z) &= J_F(x+z)^\top \nabla G(F(x+z)) \\
&= \Big(J_F(x) + F''(x)z + o(\|z\|)\Big)^\top \Big(\nabla G\left(F(x) + J_F(x)z + o(\|z\|)\right)\Big) \\
&= \Big(J_F(x) + F''(x)z + o(\|z\|)\Big)^\top \Big(\nabla G(F(x)) + \nabla^2 G(Fx)J_F(x)z + o(\|z\|)\Big) \\
&= H(x) + J_F(x)^\top \nabla^2 G(Fx)J_F(x)z + (F''(x)z)^\top \nabla G(F(x)) + o(\|z\|).
\end{aligned}$$

Note that $F''(x)z$ cannot be interpreted as a vector-matrix product since $F''(x)$ actually a 3-d tensor. In fact, it stores the Hessian for each of the components F_i of F. With this observation, we conclude. ■

Autodifferentiation

- In machine learning, many functions to deal with take the composition form

$$f(x) = F^l \circ F^{l-1} \circ \cdots \circ F^1(x).$$

- A typical example is (recurrent) neural networks :
  - let $g(u; x)$ be a fully connected multilayer neural network with input u and parameters $x = (W_1, b_1, W_2, b_2, \ldots$ W_i is the weight and b_i the bias at layer i.
  - For $\varphi : \mathbb{R} \to \mathbb{R}$ an activation map which acts componentwise on the entries of a vector, $g(u; x)$ can be defined recursively as

$$\begin{cases} g^0(u, x) & = u, \\ g^i(u, x) & = \varphi\left(W_i g^{i-1}(u, x) + b_i\right), \quad \text{for } i = 1, \ldots, p-1, \\ g(u, x) & = W_p g^{p-1}(u, x) + b_p. \end{cases}$$

  - The goal is to learn x by minimizing

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \ell(v_i, g(u_i; x)).$$

  - Denoting the affine operators $A_p = W_p \cdot + b_p$, we can also write

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \ell\Big(v_i, A_p \circ \varphi \circ A_{p-1} \circ \varphi \circ \cdots \varphi \circ A_1(u_i)\Big).$$

Autodifferentiation

- In machine learning, many functions to deal with take the composition form

$$f(x) = F^l \circ F^{l-1} \circ \cdots \circ F^1(x).$$

- A typical example is (recurrent) neural networks :
  - let $g(u;x)$ be a fully connected multilayer neural network with input u and parameters $x = (W_1, b_1, W_2, b_2, \ldots$ W_i is the weight and b_i the bias at layer i.
  - For $\varphi : \mathbb{R} \to \mathbb{R}$ an activation map which acts componentwise on the entries of a vector, $g(u;x)$ can be defined recursively as

$$\begin{cases} g^0(u,x) & = u, \\ g^i(u,x) & = \varphi\left(W_i g^{i-1}(u,x) + b_i\right), \quad \text{for } i = 1, \ldots, p-1, \\ g(u,x) & = W_p g^{p-1}(u,x) + b_p. \end{cases}$$

  - The goal is to learn x by minimizing

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \ell(v_i, g(u_i;x)).$$

  - Denoting the affine operators $A_p = W_p \cdot + b_p$, we can also write

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \ell\Big(v_i, A_p \circ \varphi \circ A_{p-1} \circ \varphi \circ \cdots \varphi \circ A_1(u_i)\Big).$$

- Fundamental question : **how to compute efficiently differential quantities of** f (typically the gradient) by exploiting this structure.
- **Autodifferentiation** [Linnainmaa 1976, Griewank and A. Walther 2008] is about computing recursively and efficiently the chain rule (up to machine precision), by exploiting the fact that every computer calculation executes a sequence of elementary arithmetic operations and elementary functions.
- Two modes : forward and reverse, the reverse one is known as backpropagation in ML/NN literature.

Autodifferentiation

- A concrete, yet general, example actually found in all situations of ML in mind is where

$$f(x) = \ell \circ F^{l-1} \circ \cdots \circ F^0(x),$$

where ℓ is a scalar-valued function (e.g. risk in ML).

- One can compute $\nabla f(x)$ using the chain rule in two steps :

1. Forward pass : compute the function value and keep track of all the intermediate computations,

$$\beta_0 = x, \quad \beta_{i+1} \overset{\text{def}}{=} F^i(\beta_i), \quad f(x) = \ell(\beta_l).$$

2. Backward pass : compute the chain rule $\nabla f(x) = J_{F^0}(\beta_0)^\top \cdots J_{F^{l-1}}(\beta_{l-1})^\top \nabla \ell(\beta_l)$ with backward recursion

$$h_l = \nabla \ell(\beta_l), \quad h_{i-1} \overset{\text{def}}{=} J_{F^{i-1}}(\beta_{i-1})^\top h_i.$$

Autodifferentiation

- A concrete, yet general, example actually found in all situations of ML in mind is where

$$f(x) = \ell \circ F^{l-1} \circ \cdots \circ F^0(x),$$

where ℓ is a scalar-valued function (e.g. risk in ML).

- One can compute $\nabla f(x)$ using the chain rule in two steps :

1. Forward pass : compute the function value and keep track of all the intermediate computations,

$$\beta_0 = x, \quad \beta_{i+1} \overset{\text{def}}{=} F^i(\beta_i), \quad f(x) = \ell(\beta_l).$$

2. Backward pass : compute the chain rule $\nabla f(x) = J_{F^0}(\beta_0)^\top \cdots J_{F^{l-1}}(\beta_{l-1})^\top \nabla \ell(\beta_l)$ with backward recursion

$$h_l = \nabla \ell(\beta_l), \quad h_{i-1} \overset{\text{def}}{=} J_{F^{i-1}}(\beta_{i-1})^\top h_i.$$

- The backward step entails vector-matrix multiplications (a forward accumulation would be even more prohibitive with matrix-matrix multiplications).
- The main issue is that for ML :
  - these transpose Jacobian matrices are difficult to apply ;
  - it is out of question to store them on a computer ;
  - entails quadratic complexity in space and time.
- Autodifferentiation is about differentiating automatically any function which can be implemented on a computer with **the same computational cost as for evaluating the function itself**.
- Autodifferentiation is a cornerstone of modern data science.

Function value evaluation

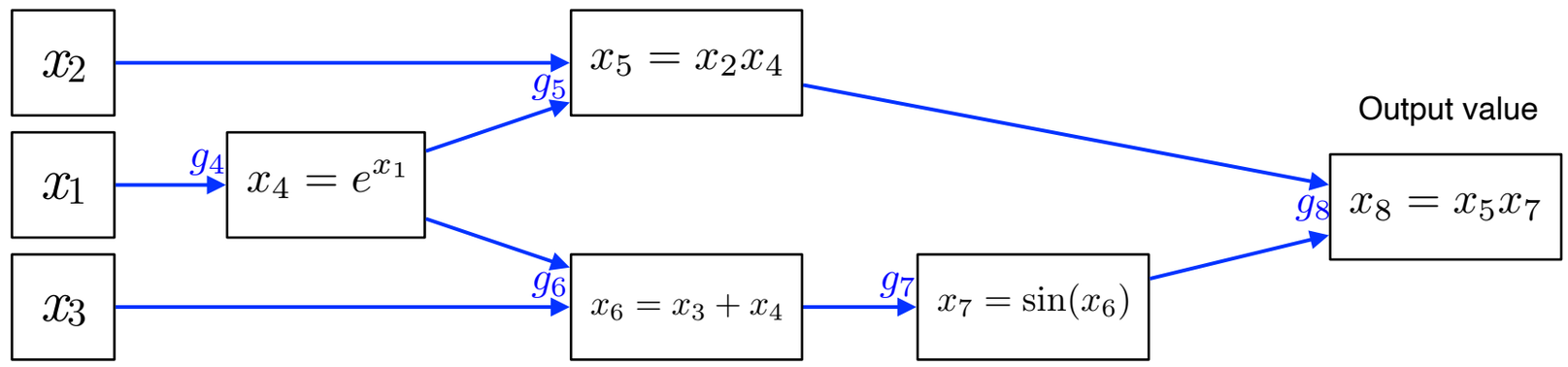
- The key idea is that the *computational graph* of any computable function f can be represented as a directed acyclic graph (DAG).
- $(x_i)_{i\in[r]}$ the set of all scalar variables (input, output and intermediate) manipulated by the computational graph :
  - $(x_1, \cdots, x_d)$ are the input variables.
  - $(x_{d+1}, \cdots, x_{r-1})$ are the intermediate variables.
  - $x_r = f(x_1, \cdots, x_{r-1})$ is the output variable.

Function value evaluation

- The key idea is that the *computational graph* of any computable function f can be represented as a directed acyclic graph (DAG).
- $(x_i)_{i\in[r]}$ the set of all scalar variables (input, output and intermediate) manipulated by the computational graph :
  - $(x_1, \cdots, x_d)$ are the input variables.
  - $(x_{d+1}, \cdots, x_{r-1})$ are the intermediate variables.
  - $x_r = f(x_1, \cdots, x_{r-1})$ is the output variable.

$$f(x_1, x_2, x_3) = x_2 e^{x_1} \sin(x_3 + e^{x_1})$$

Computational graph as a DAG

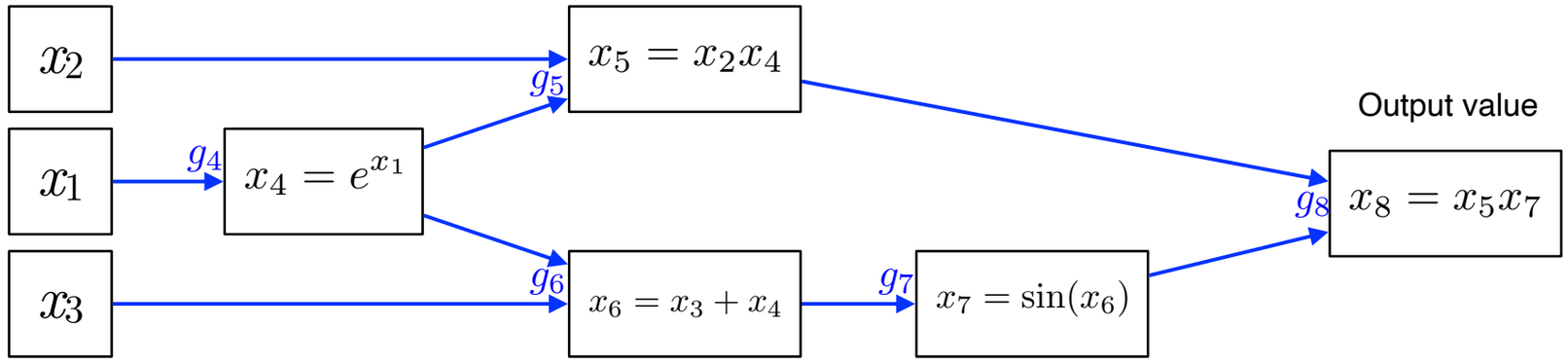


Function value evaluation

- The key idea is that the *computational graph* of any computable function f can be represented as a directed acyclic graph (DAG).
- $(x_i)_{i\in[r]}$ the set of all scalar variables (input, output and intermediate) manipulated by the computational graph :
  - $(x_1, \cdots, x_d)$ are the input variables.
  - $(x_{d+1}, \cdots, x_{r-1})$ are the intermediate variables.
  - $x_r = f(x_1, \cdots, x_{r-1})$ is the output variable.

$$f(x_1, x_2, x_3) = x_2 e^{x_1} \sin(x_3 + e^{x_1})$$

Computational graph as a DAG



- Function value evaluation: forward pass as a DAG traversal:

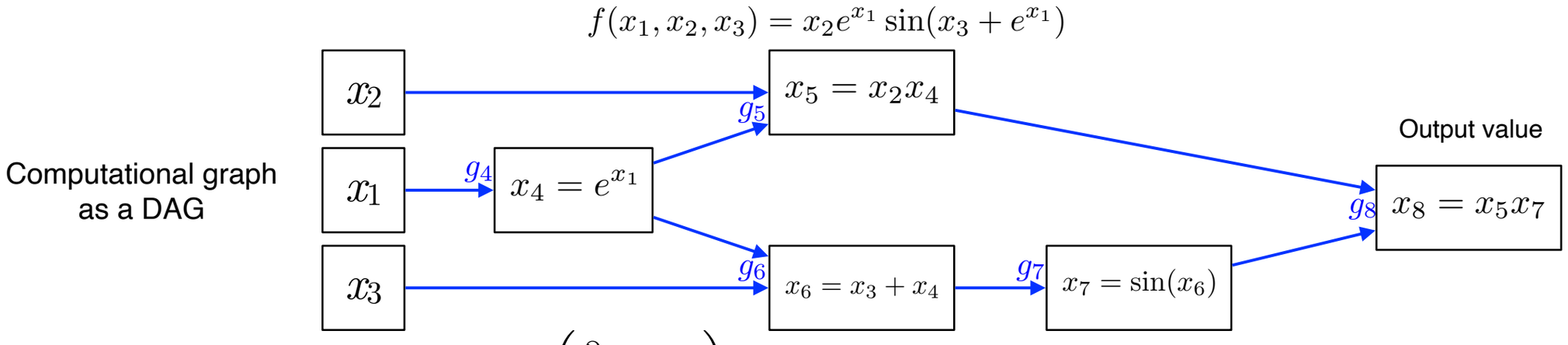
Input : values of $(x_1, \cdots, x_d)$; scalar-valued elementary functions $g_i : \mathbb{R}^{|\texttt{parents}(i)|} \to \mathbb{R}$, $d+1 \leq i \leq r$;

for $i = d+1$ *to* r **do**

  $x_i = g_i(x_{\texttt{parents}(i)})$; where $x_{\texttt{parents}(i)} = (x_j)_{j\in\texttt{parents}(i)}$, and $\texttt{parents}(i) \subseteq [r-1]$ is the set of parent nodes of i in the DAG.

return x_r.

Forward mode autodifferentiation



- The goal is to compute $\nabla f(x) = \left(\frac{\partial x_r}{x_j}(x)\right)_{j \in [d]}$.
- Apply the "forward" chain rule : for $j = 1, \ldots, d$, compute the partial derivatives

$$\frac{\partial x_i}{\partial x_j} = \sum_{k \in \texttt{parents}(i)} \frac{\partial x_k}{\partial x_j} \frac{\partial g_i}{\partial x_k}.$$

- Accumulate by forward pass (DAG traversal).

Input : values of $(x_1, \cdots, x_d)$;

Initialization : $\frac{\partial x_j}{\partial x_j} = 1$;

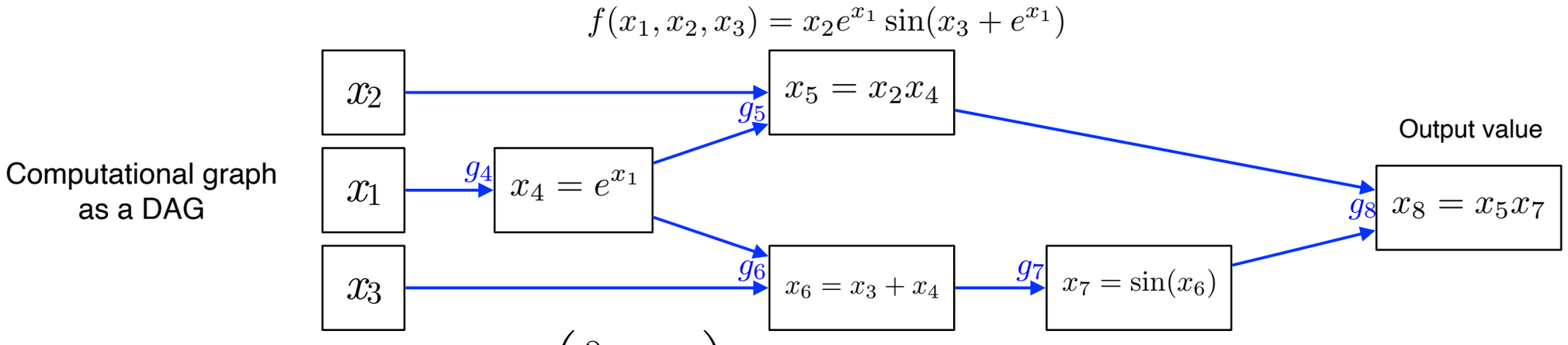
for $i = d + 1$ *to* r **do**

  for $j = 1$ *to* d **do**

    $\frac{\partial x_i}{\partial x_j} = \sum_{k \in \texttt{parents}(i)} \frac{\partial x_k}{\partial x_j} \frac{\partial x_i}{\partial x_k}$.

return ∇x_r.

Forward mode autodifferentiation



- The goal is to compute $\nabla f(x) = \left(\frac{\partial x_r}{x_j}(x)\right)_{j\in[d]}$.
- Apply the "forward" chain rule : for $j = 1, \ldots, d$, compute the partial derivatives

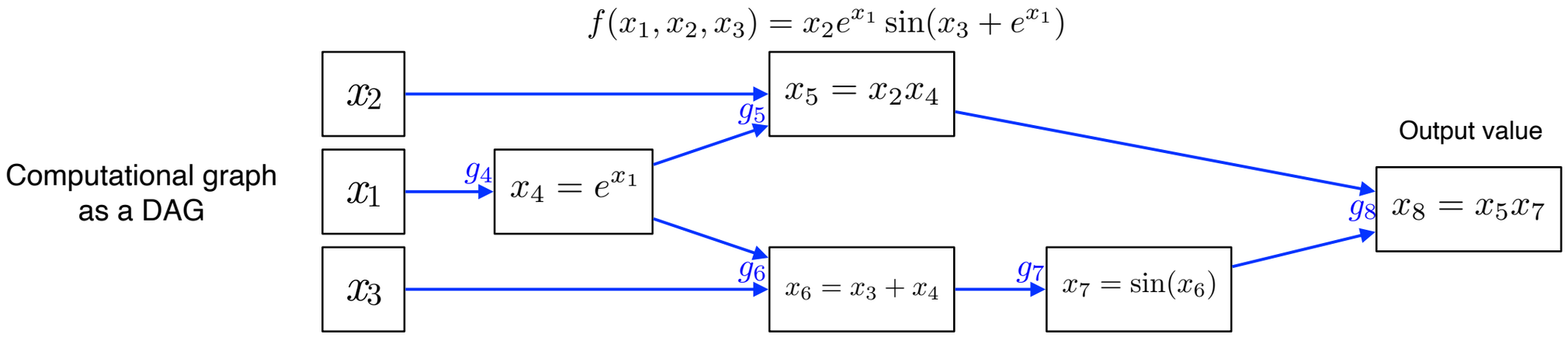
$$\frac{\partial x_i}{\partial x_j} = \sum_{k \in \texttt{parents}(i)} \frac{\partial x_k}{\partial x_j} \frac{\partial g_i}{\partial x_k}.$$

- Accumulate by forward pass (DAG traversal).

Input : values of $(x_1, \cdots, x_d)$;
Initialization : $\frac{\partial x_j}{\partial x_j} = 1$;
for $i = d+1$ *to* r **do**
  for $j = 1$ *to* d **do**
    $\frac{\partial x_i}{\partial x_j} = \sum_{k\in\texttt{parents}(i)} \frac{\partial x_k}{\partial x_j} \frac{\partial x_i}{\partial x_k}$.
return ∇x_r.

While natural, sub-optimal
Complexity d times function evaluation/DAG traversal

Reverse mode autodifferentiation

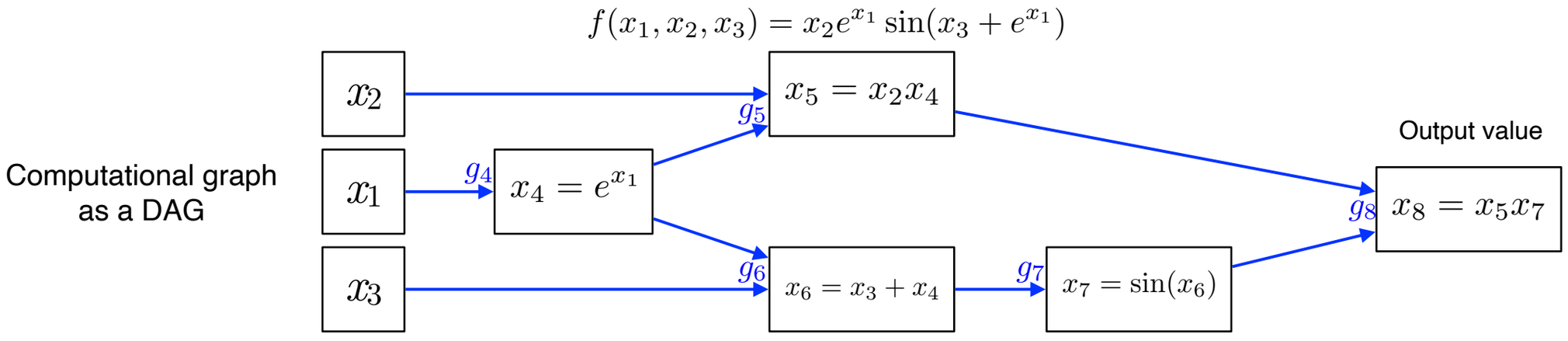


- The goal is to compute $\nabla f(x) = \left(\frac{\partial x_r}{x_j}(x)\right)_{j \in [d]}$.
- Apply the "backward" chain rule : for $i = 1, \ldots, r-1$, compute the partial derivatives

$$\frac{\partial x_r}{\partial x_i} = \sum_{k \in \texttt{childs}(i)} \frac{\partial x_r}{\partial x_k} \frac{\partial g_k}{\partial x_i}.$$

- Corresponds to a backward accumulation (reverse traversal).

Reverse mode autodifferentiation



- The goal is to compute $\nabla f(x) = \left(\frac{\partial x_r}{x_j}(x)\right)_{j\in[d]}$.
- Apply the "backward" chain rule : for $i = 1, \ldots, r-1$, compute the partial derivatives

$$\frac{\partial x_r}{\partial x_i} = \sum_{k \in \texttt{childs}(i)} \frac{\partial x_r}{\partial x_k} \frac{\partial g_k}{\partial x_i}.$$

- Corresponds to a backward accumulation (reverse traversal).

Input : values of $(x_1, \cdots, x_d)$;

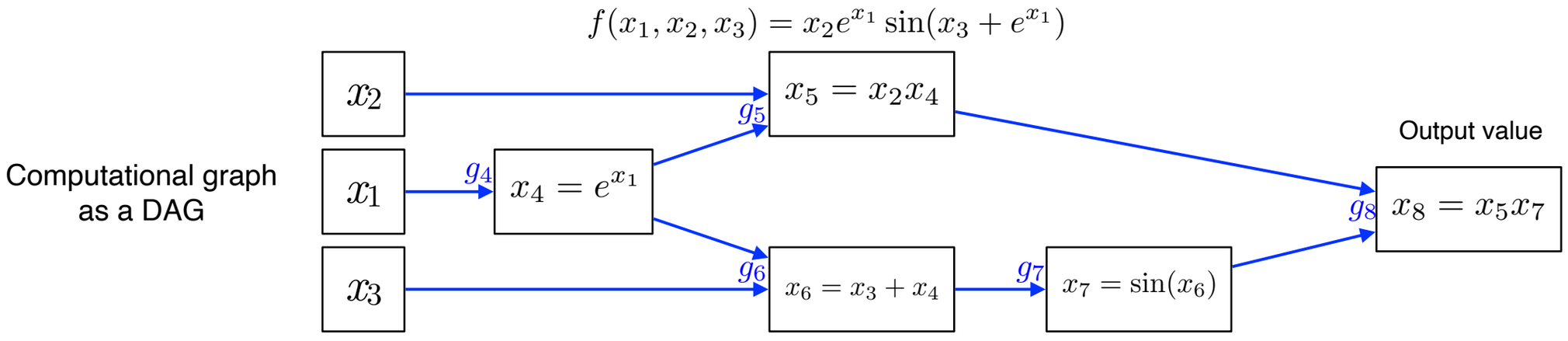
Initialization : $\nabla f = (0, \ldots, 0, 1) \in \mathbb{R}^r$;

for $i = r-1$ *to* 1 **do**

  $\nabla_i f = \sum_{k \in \texttt{childs}(i)} \nabla_k f \frac{\partial g_k}{\partial x_i}$.

return $(\nabla_1 f, \nabla_2 f, \cdots, \nabla_d f)$.

Reverse mode autodifferentiation



- The goal is to compute $\nabla f(x) = \left(\frac{\partial x_r}{x_j}(x)\right)_{j\in[d]}$.
- Apply the "backward" chain rule : for $i = 1, \ldots, r-1$, compute the partial derivatives

$$\frac{\partial x_r}{\partial x_i} = \sum_{k \in \texttt{childs}(i)} \frac{\partial x_r}{\partial x_k} \frac{\partial g_k}{\partial x_i}.$$

- Corresponds to a backward accumulation (reverse traversal).

Input : values of $(x_1, \cdots, x_d)$;

Initialization : $\nabla f = (0, \ldots, 0, 1) \in \mathbb{R}^r$;

for $i = r-1$ *to* 1 **do**

  $\nabla_i f = \sum_{k \in \texttt{childs}(i)} \nabla_k f \frac{\partial g_k}{\partial x_i}$.

return $(\nabla_1 f, \nabla_2 f, \cdots, \nabla_d f)$.

Much better: only one (backward) pass

Theorem: Same complexity as function evaluation.

Demo Autograd

Gradient-Lipschitz functions

Definition *A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is* L*-smooth if and only if it is differentiable and its gradient is* L*-Lipschitz-continuous, i.e.*

$$\|\nabla f(x) - \nabla f(z)\| \leq L \|x - z\|, \quad \forall x, z, \in \mathbb{R}^d.$$

We then say that $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$.

Gradient-Lipschitz functions

Definition *A function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if and only if it is differentiable and its gradient is L-Lipschitz-continuous, i.e.*

$$\|\nabla f(x) - \nabla f(z)\| \leq L \|x - z\|, \quad \forall x, z, \in \mathbb{R}^d.$$

We then say that $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$.

Proposition *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable. Then $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ if and only if $\nabla^2 f(x) \preceq L \cdot \mathrm{I}$ for all $x \in \mathbb{R}^d$, i.e.,*

$$\left\langle y, \nabla^2 f(x) y \right\rangle \leq L \|y\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

Gradient-Lipschitz functions

Definition *A function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if and only if it is differentiable and its gradient is L-Lipschitz-continuous, i.e.*

$$\|\nabla f(x) - \nabla f(z)\| \leq L \|x - z\|, \quad \forall x, z, \in \mathbb{R}^d.$$

We then say that $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$.

Proposition *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable. Then $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ if and only if $\nabla^2 f(x) \preceq L \cdot \mathrm{I}$ for all $x \in \mathbb{R}^d$, i.e.,*

$$\left\langle y, \nabla^2 f(x) y \right\rangle \leq L \|y\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

Example (Machine learning)

- $f(x) = \frac{1}{n} \sum_{i=1}^n \ell(v_i, x^\top \varphi(u_i))$.
- *If $\ell \in \mathscr{C}_{L_\ell}^{1,1}(\mathbb{R})$ and D-bounded data, then $f \in \mathscr{C}_{L_\ell D^2}^{1,1}(\mathbb{R}^d)$ (check by computing the gradient).*
- *If ℓ is also twice-differentiable, then (see S32)*

$$\nabla^2 f(x) = \frac{1}{n} \sum_{i=1}^n \varphi(u_i) \ell''(v_i, x^\top \varphi(u_i)) \varphi(u_i)^\top \approx L_\ell \underbrace{\frac{1}{n} \sum_{i=1}^n \varphi(u_i) \varphi(u_i)^\top}_{\text{Covariance matrix (Empirical)}}.$$

Gradient-Lipschitz functions

Proof: $\Leftarrow$ Suppose that $f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d)$. Define the function $\phi_y(x) = \langle y, \nabla f(x)\rangle$. By the Cauchy-Schwartz-inequality, we have $\phi_y \in \mathscr{C}^{1,1}_{L\|y\|}(\mathbb{R}^d)$. Thus,

$$|\phi_y(x+ty) - \phi_y(x)| \leq Lt\|y\|^2. \tag{1}$$

Moreover, by linearity

$$\lim_{t\to 0} \frac{\phi_y(x+ty)-\phi_y(x)}{t} = \lim_{t\to 0} \frac{\langle y, (\nabla f(x+ty) - \nabla f(x))\rangle}{t} = \langle y, \nabla^2 f(x) y\rangle.$$

Using this after passing to the limit as $t \to 0$ in (1), we conclude.

$\Rightarrow$ Suppose that $\nabla^2 f(x) \preceq L \cdot \mathrm{I}$. Then, by Taylor expansion with an integral remainder, we have

$$\nabla f(x) - \nabla f(z) = \int_0^1 \nabla^2 f(z + t(x-z))(x-z)dt.$$

Therefore

$$\begin{aligned}\|\nabla f(x) - \nabla f(z)\| &\leq \int_0^1 \left\|\nabla^2 f(z+t(x-z))(x-z)\right\| \leq \|x-z\| \int_0^1 \left|\!\left|\!\left|\nabla^2 f(z+t(x-z))\right|\!\right|\!\right| dt \\ &= \|x-z\| \int_0^1 \sup_{y\in\mathbb{R}^d} \frac{\langle y, \nabla^2 f(z+t(x-z))y\rangle}{\|y\|^2} dt \leq L\|x-z\|.\end{aligned}$$

■

Descent lemma

Lemma *If* $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$, *then* $\forall x, z \in \mathbb{R}^d$,

$$|f(z) - f(x) - \langle \nabla f(x), z - x \rangle| \leq \frac{L}{2} \|z - x\|^2 .$$

Clearly, f can be well approximated locally by a quadratic function.

Descent lemma

Lemma *If* $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$, *then* $\forall x, z \in \mathbb{R}^d$,

$$|f(z) - f(x) - \langle \nabla f(x), z - x \rangle| \leq \frac{L}{2} \|z - x\|^2 .$$

Clearly, f can be well approximated locally by a quadratic function.

The terminology "descent lemma" will be clear when applied to algorithms

Descent lemma

Lemma *If* $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$, *then* $\forall x, z \in \mathbb{R}^d$,

$$|f(z) - f(x) - \langle \nabla f(x), z - x \rangle| \leq \frac{L}{2} \|z - x\|^2 .$$

Clearly, f can be well approximated locally by a quadratic function.

The terminology "descent lemma" will be clear when applied to algorithms

Proof: By Taylor expansion with an integral remainder, we have

$$f(z) - f(x) = \int_0^1 \langle \nabla f(x + t(z - x)), z - x \rangle dt.$$

Thus,

$$\begin{aligned}
|f(z) - f(x) - \langle \nabla f(x), z - x \rangle| &= \left| \int_0^1 \langle \nabla f(x + t(z - x)) - \nabla f(x), z - x \rangle dt \right| \\
\text{(Cauchy-Schwartz)} \quad &\leq \|z - x\| \left(\int_0^1 \|\nabla f(x + t(z - x)) - \nabla f(x)\| \, dt \right) \\
\text{(Lipschitz continuity of the gradient)} \quad &\leq \|z - x\| \left(\int_0^1 Lt \|z - x\| \, dt \right) \\
&\leq \frac{L}{2} \|z - x\|^2 .
\end{aligned}$$

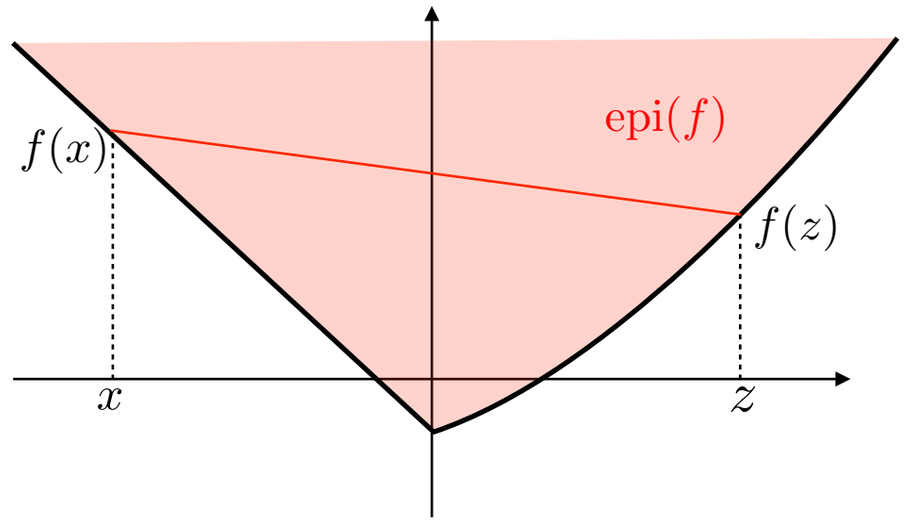
Convexity

Global definition without differentiability

Definition *A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is convex iff* $\forall x, z \in \mathbb{R}^d$, $\forall \rho \in]0,1[$

$$f(\rho x + (1-\rho)z) \leq \rho f(x) + (1-\rho) f(z).$$

If the inequality is strict for $x \neq z$, f *is strictly convex.*



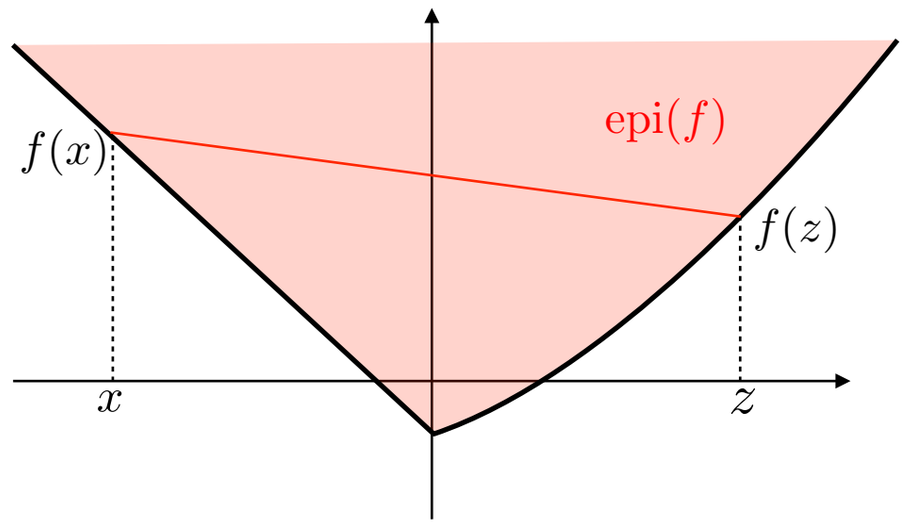
Convexity

Global definition without differentiability

Definition *A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is convex iff* $\forall x, z \in \mathbb{R}^d$, $\forall \rho \in]0, 1[$

$$f(\rho x + (1-\rho)z) \leq \rho f(x) + (1-\rho) f(z).$$

If the inequality is strict for $x \neq z$, f *is strictly convex.*

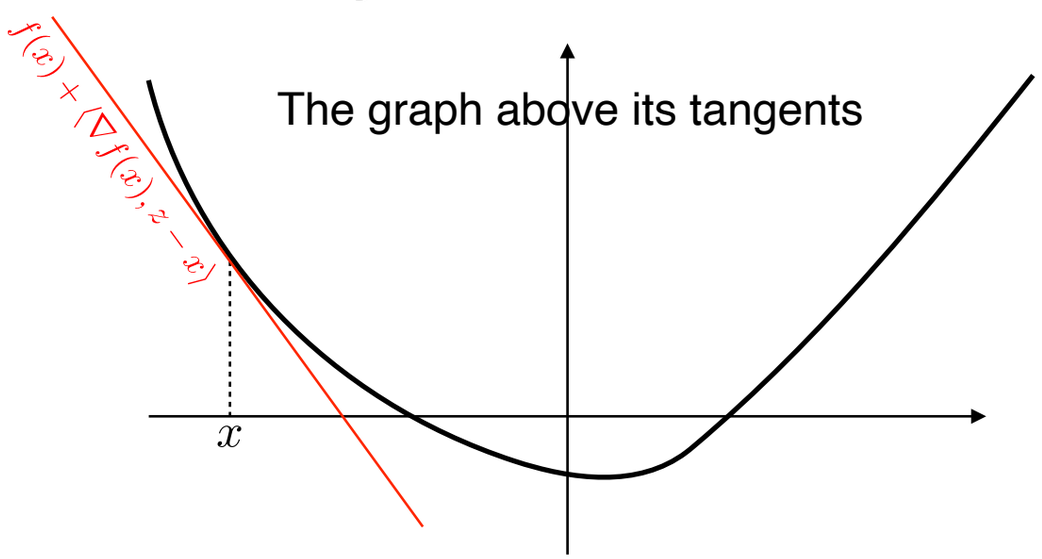


Global definition with differentiability

Definition *A differentiable function* $f : \mathbb{R}^d \to \mathbb{R}$ *is convex iff* $\forall x, z \in \mathbb{R}^d$

$$f(z) \geq f(x) + \langle \nabla f(x), z - x \rangle.$$

If the inequality is strict for $x \neq z$, f *is strictly convex.*



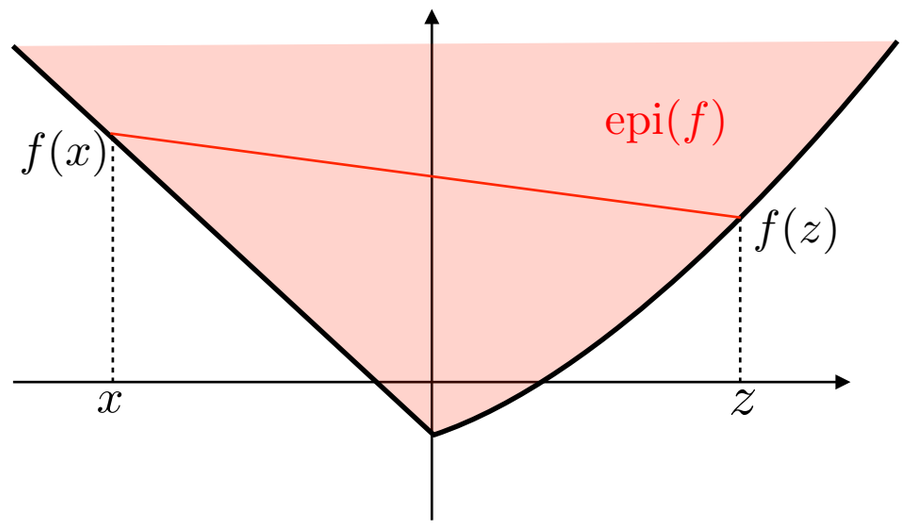
Convexity

Global definition without differentiability

Definition *A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is convex iff* $\forall x, z \in \mathbb{R}^d$, $\forall \rho \in]0, 1[$

$$f(\rho x + (1-\rho)z) \leq \rho f(x) + (1-\rho) f(z).$$

If the inequality is strict for $x \neq z$, f *is strictly convex.*

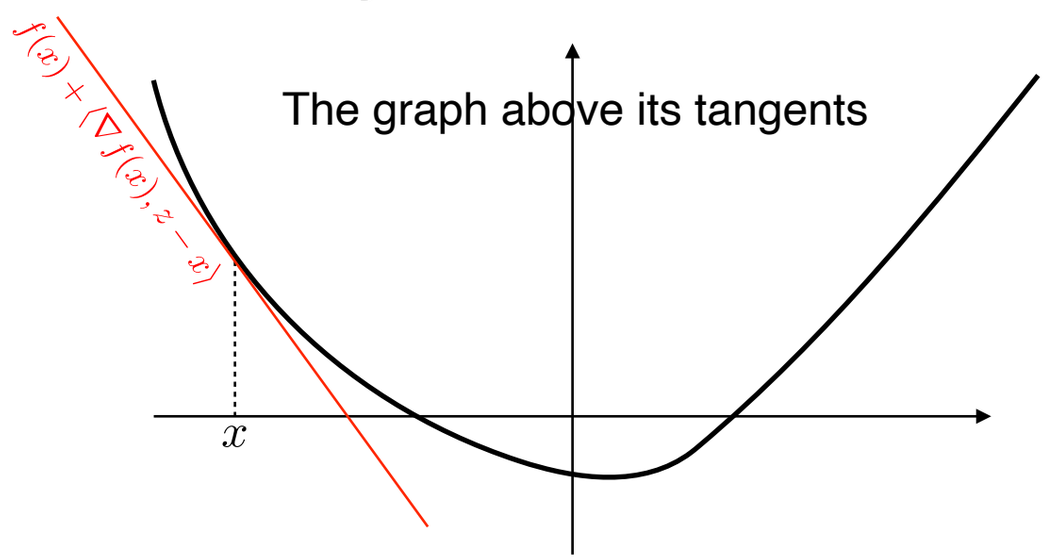


Global definition with differentiability

Definition *A differentiable function* $f : \mathbb{R}^d \to \mathbb{R}$ *is convex iff* $\forall x, z \in \mathbb{R}^d$

$$f(z) \geq f(x) + \langle \nabla f(x), z - x \rangle.$$

If the inequality is strict for $x \neq z$, f *is strictly convex.*



Local definition with differentiability

Definition *A twice differentiable function* $f : \mathbb{R}^d \to \mathbb{R}$ *is convex iff* $\nabla^2 f(x) \succeq 0$, $\forall x \in \mathbb{R}^d$.

The graph has non-negative curvature

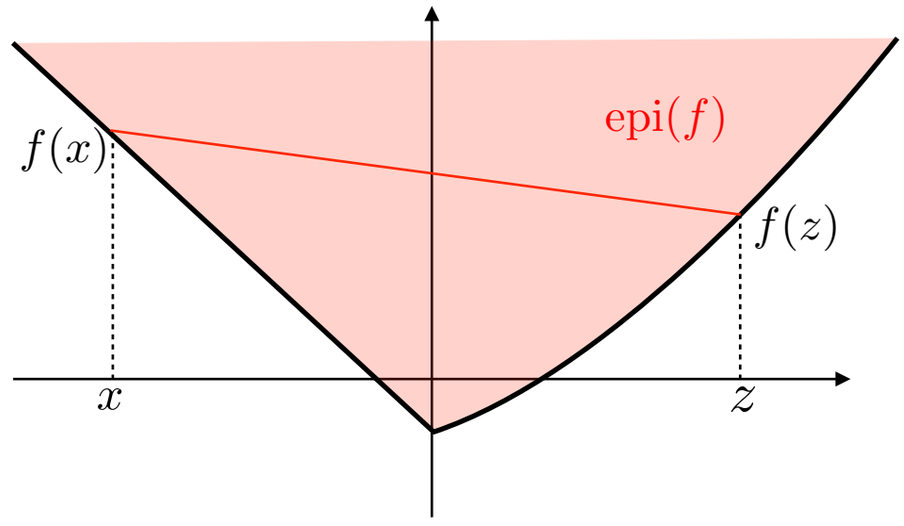
Convexity

Global definition without differentiability

Definition *A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is convex iff* $\forall x, z \in \mathbb{R}^d$, $\forall \rho \in]0,1[$

$$f(\rho x + (1-\rho)z) \leq \rho f(x) + (1-\rho) f(z).$$

If the inequality is strict for $x \neq z$, f *is strictly convex.*



Lemma (Jensen's inequality) *Let* f *be convex. Then for any points* $x_1, \ldots, x_n \in \mathbb{R}^d$, *and scalars* $(\rho_1, \ldots, \rho_n) \in [0, +\infty[^n$ *such that* $\sum_{i=1}^n \rho = 1$, *it holds*

$$f\left(\sum_{i=1}^n \rho_i x_i\right) \leq \sum_{i=1}^n \rho_i f(x_i).$$

More generally, if X *is a random variable, then*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

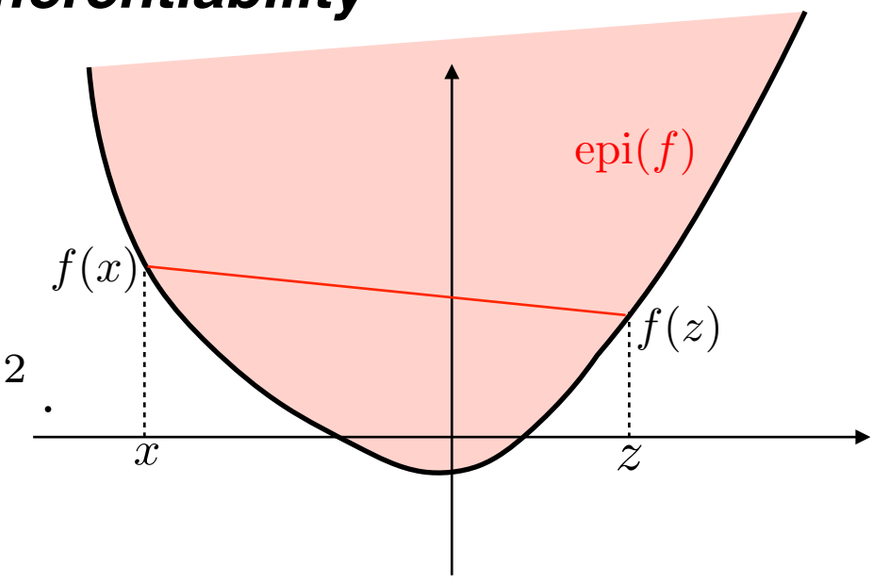
Proof: The first inequality follows by induction and the convexity inequality. The probabilistic version of the inequality can be proved using the monotonicity of the (sub)derivative. ■

Strong convexity

Global definition without differentiability

Definition *A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is* μ*-strongly convex,* $\mu > 0$*, iff* $\forall x, z \in \mathbb{R}^d$, $\forall \rho \in]0,1[$

$$f(\rho x + (1-\rho)z) \leq \rho f(x) + (1-\rho) f(z) - \rho(1-\rho)\frac{\mu}{2}\|x - z\|^2 .$$

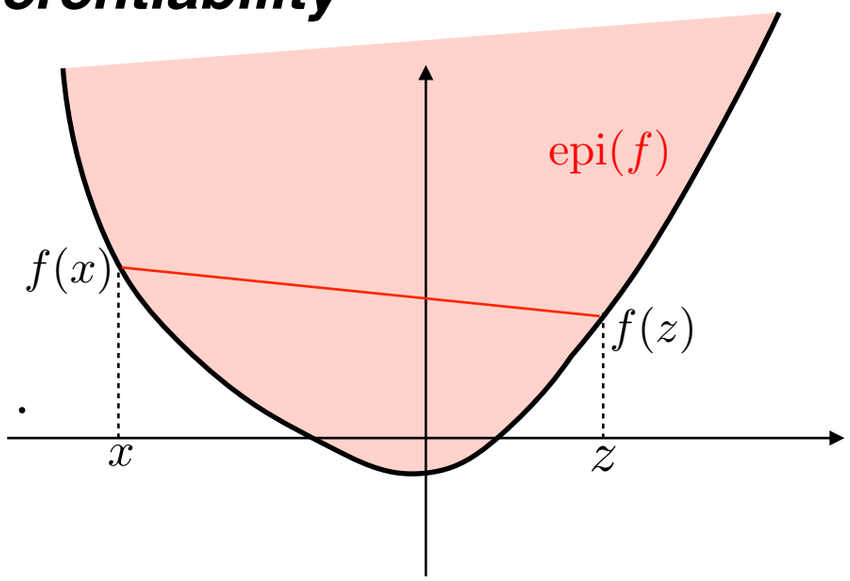


Strong convexity

Global definition without differentiability

Definition *A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is* μ*-strongly convex,* $\mu > 0$*, iff* $\forall x, z \in \mathbb{R}^d$, $\forall \rho \in]0,1[$

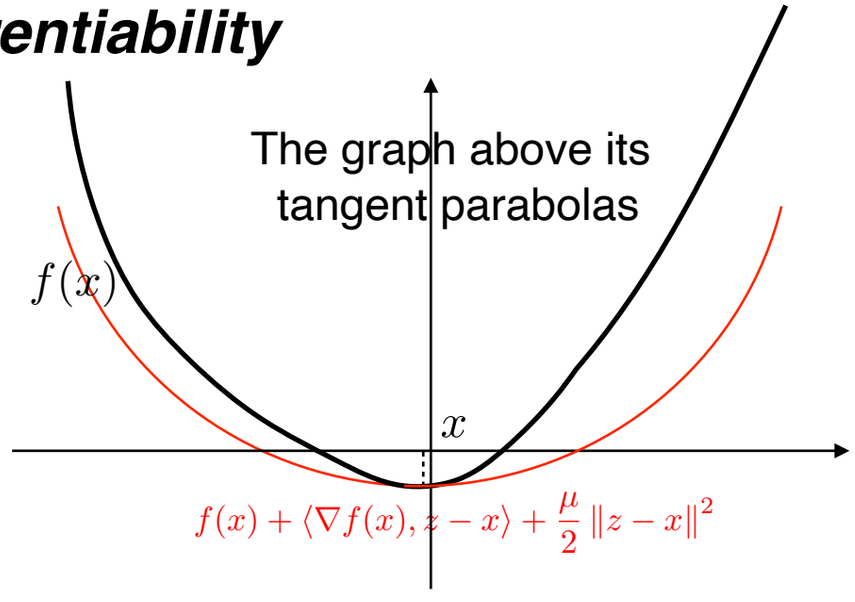
$$f(\rho x + (1-\rho)z) \leq \rho f(x) + (1-\rho) f(z) - \rho(1-\rho)\frac{\mu}{2}\|x - z\|^2 .$$



Global definition with differentiability

Definition *A differentiable function* $f : \mathbb{R}^d \to \mathbb{R}$ *is* μ*-strongly convex iff* $\forall x, z \in \mathbb{R}^d$

$$f(z) \geq f(x) + \langle \nabla f(x), z - x \rangle + \frac{\mu}{2}\|z - x\|^2 .$$

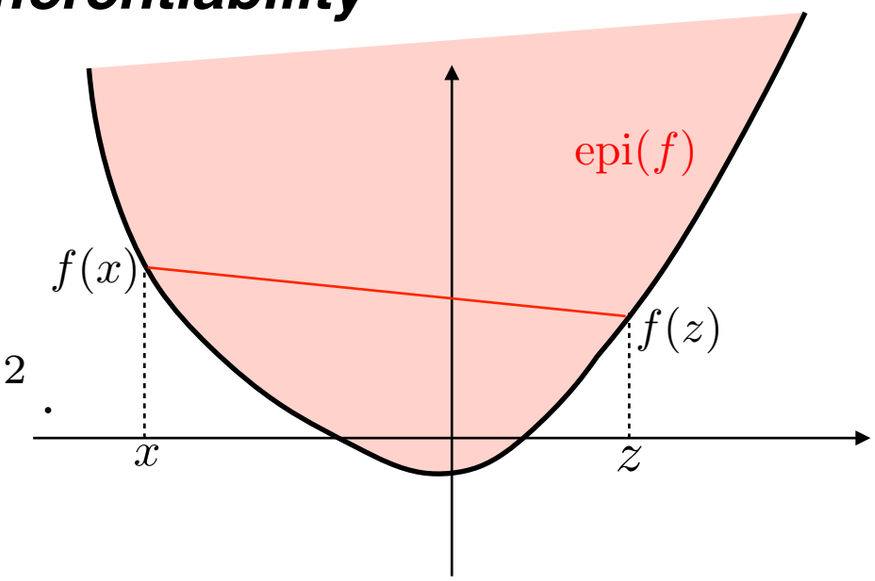


Strong convexity

Global definition without differentiability

Definition *A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is* μ*-strongly convex,* $\mu > 0$*, iff* $\forall x, z \in \mathbb{R}^d$, $\forall \rho \in]0,1[$

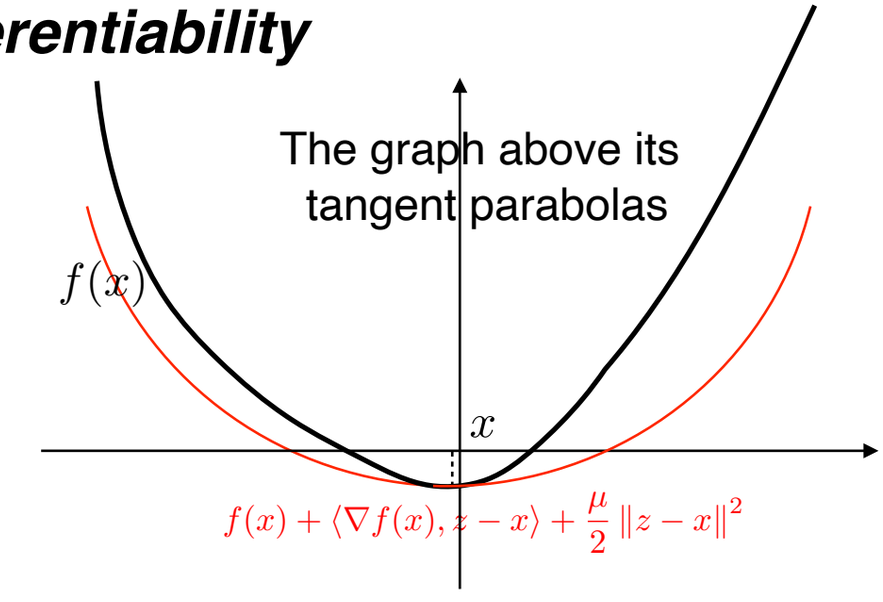
$$f(\rho x + (1-\rho)z) \leq \rho f(x) + (1-\rho) f(z) - \rho(1-\rho)\frac{\mu}{2}\|x - z\|^2 .$$



Global definition with differentiability

Definition *A differentiable function* $f : \mathbb{R}^d \to \mathbb{R}$ *is* μ*-strongly convex iff* $\forall x, z \in \mathbb{R}^d$

$$f(z) \geq f(x) + \langle \nabla f(x), z - x \rangle + \frac{\mu}{2}\|z - x\|^2 .$$



Local definition with differentiability

Definition *A twice differentiable function* $f : \mathbb{R}^d \to \mathbb{R}$ *is* μ*-strongly convex iff* $\nabla^2 f(x) \succeq \mu \cdot \mathrm{I}, \forall x \in \mathbb{R}^d$.

The graph is positively curved

Strong convexity

Example (Machine learning)

- $f(x) = \frac{1}{n}\sum_{i=1}^{n} \ell(v_i, x^\top \varphi(u_i))$.
- *If ℓ is differentiable and μ_ℓ-strongly convex, then for all x, z and $i \in [n]$,*

$$\begin{aligned}\ell(v_i, z^\top \varphi(u_i)) &\geq \ell(v_i, x^\top \varphi(u_i)) + \ell'(v_i, x^\top \varphi(u_i))(z-x)^\top \varphi(u_i) + \frac{\mu_\ell}{2}|(z-x)^\top \varphi(u_i)|^2 \\ &= \ell(v_i, x^\top \varphi(u_i)) + \ell'(v_i, x^\top \varphi(u_i))(z-x)^\top \varphi(u_i) + \frac{\mu_\ell}{2}\big\langle z-x, \varphi(u_i)\varphi(u_i)^\top (z-x)\big\rangle.\end{aligned}$$

Averaging over i and the chain rule to see that

$$\nabla f(x) = \frac{1}{n}\sum_{i=1}^{n} \ell'(v_i, x^\top \varphi(u_i))\varphi(u_i),$$

we have

$$f(z) \geq f(x) + \langle \nabla f(x), z-x\rangle + \frac{\mu_\ell}{2}\left\langle z-x, \left(\frac{1}{n}\sum_{i=1}^{n} \varphi(u_i)\varphi(u_i)^\top\right)(z-x)\right\rangle.$$

- *f is strongly convex iff the covariance matrix $\frac{1}{n}\sum_{i=1}^{n} \varphi(u_i)\varphi(u_i)^\top$ is invertible (low correlation/dimension).*
- *If lack of strong convexity, add $\frac{\mu}{2}\|\cdot\|^2$ to f : be careful with the choice of μ to avoid additional bias.*

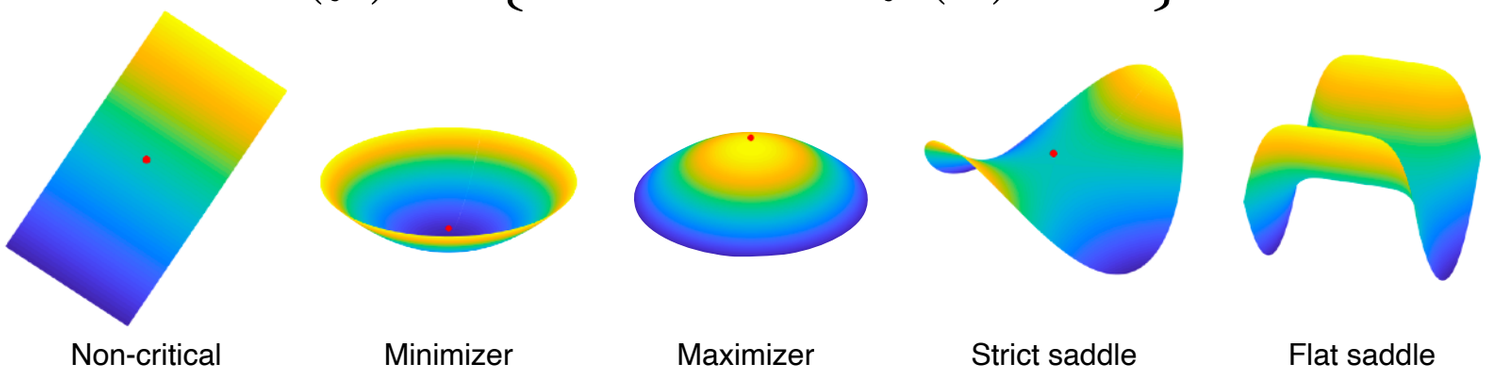
Why (strong) convexity

- Convexity is preserved under many operators : e.g.
  - positive sum,
  - post-composition by an affine operator,
  - $\max$ of a family of convex functions is convex.
- Convexity allows for duality theory.

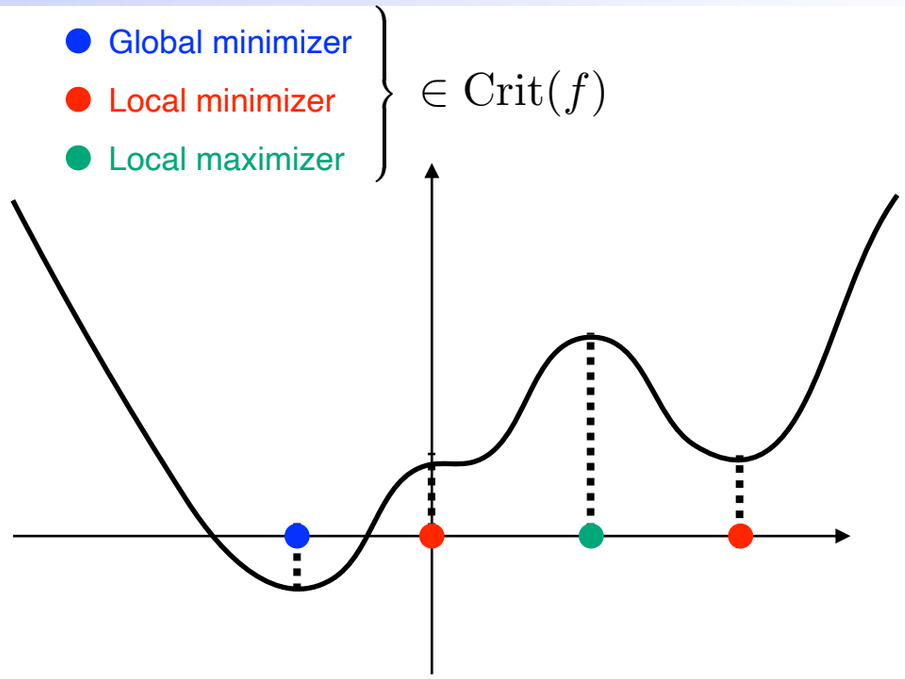
Fermat's rule and role of convexity

Definition *The set of critical points of a differentiable function f is*

$$\mathrm{Crit}(f) = \left\{x \in \mathbb{R}^d : \ \nabla f(x) = 0\right\}.$$



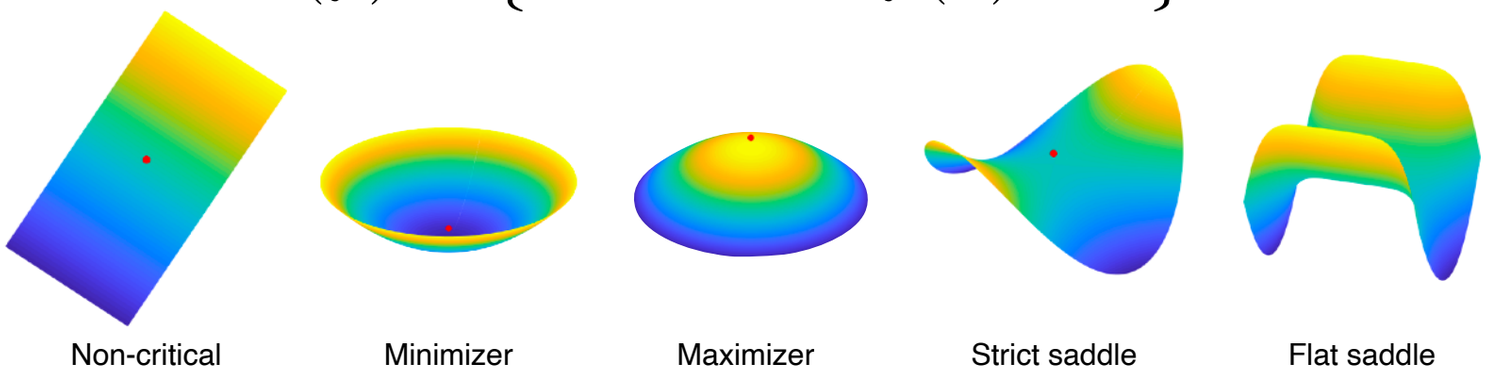
Non-critical | Minimizer | Maximizer | Strict saddle | Flat saddle

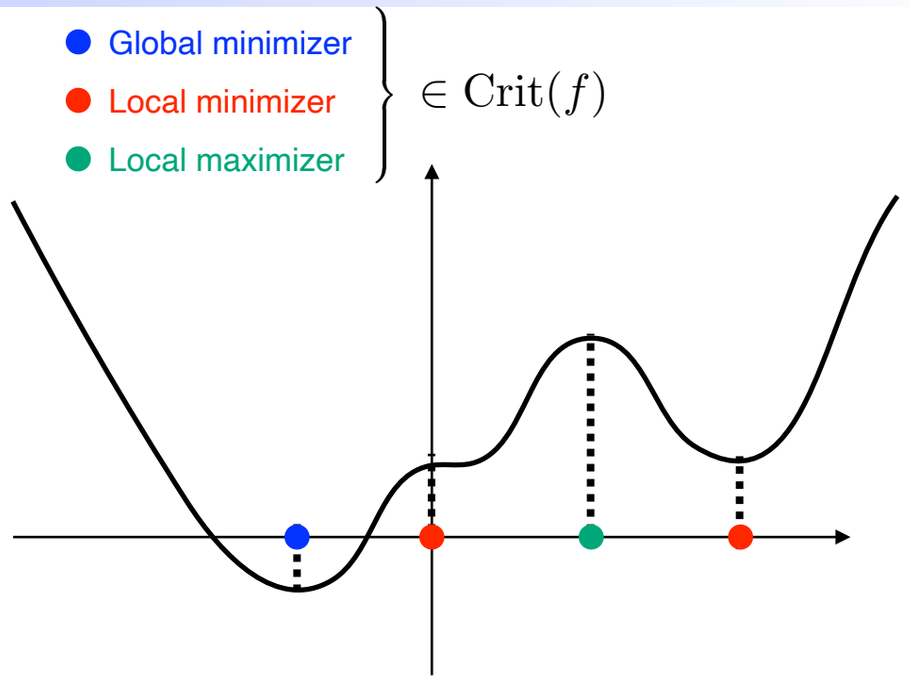


Fermat's rule and role of convexity

Definition *The set of critical points of a differentiable function f is*

$$\mathrm{Crit}(f) = \left\{x \in \mathbb{R}^d : \ \nabla f(x) = 0\right\}.$$



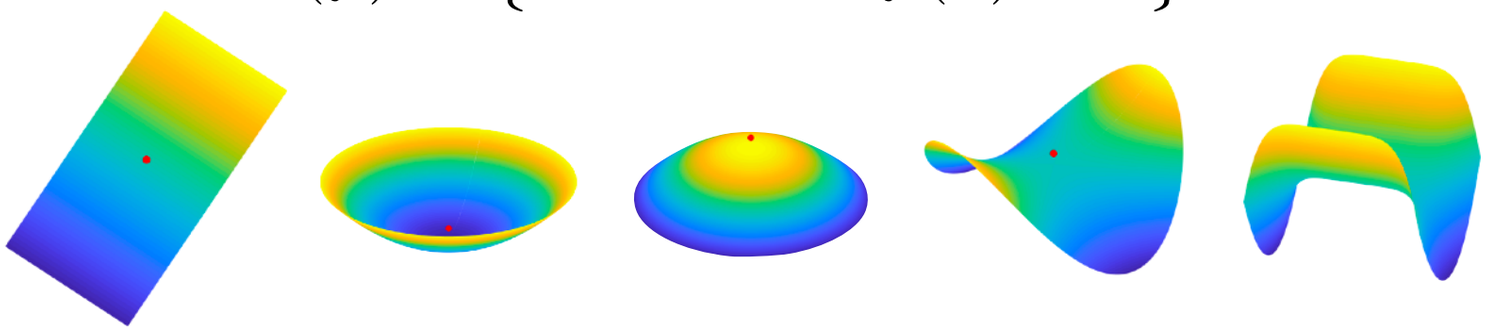


Theorem *If $x^\star$ is a (local) minimizer of a smooth differentiable function f, then $\nabla f(x^\star) = 0$.*

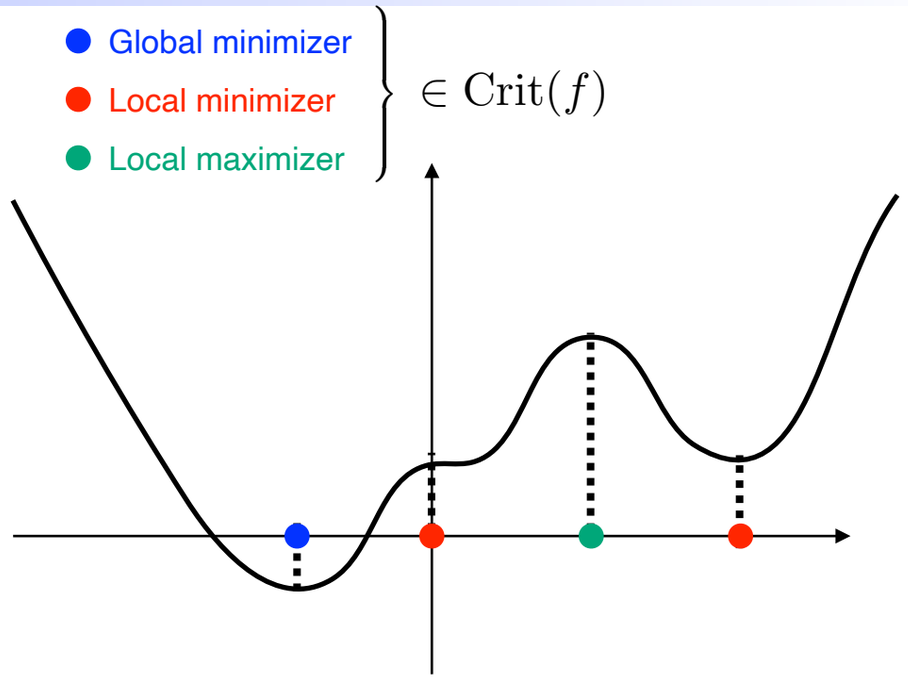
Fermat's rule and role of convexity

Definition *The set of critical points of a differentiable function f is*

$$\mathrm{Crit}(f) = \left\{x \in \mathbb{R}^d : \ \nabla f(x) = 0\right\}.$$



Non-critical | Minimizer | Maximizer | Strict saddle | Flat saddle



Theorem *If $x^\star$ is a (local) minimizer of a smooth differentiable function f, then $\nabla f(x^\star) = 0$.*

Proof: If $x^\star$ is a local minimizer, there exists $\epsilon > 0$ such that

$$f(x^\star + z) \geq f(x^\star), \quad \forall z \in \mathbb{B}_\epsilon(0).$$

By Taylor formula, we get $\forall z \in \mathbb{B}_\epsilon(0)$

$$\langle -\nabla f(x^\star), z\rangle \leq o(\|z\|)$$

or, for any unit norm vector z,

$$\langle -\nabla f(x^\star), z\rangle \leq o(1).$$

This shows that $\langle -\nabla f(x^\star), z\rangle = 0$, and since z is arbitrary unit norm vector, we have necessarily $\nabla f(x^\star) = 0$. ■

Role of convexity in minimization

- In general, $\underset{\mathbb{R}^d}{\mathrm{Argmin}}\,(f) \subset \mathrm{Crit}(f)$.
- f convex : all critical points are global minimizers (when they exist), i.e.

$$\mathrm{Crit}(f) = \underset{\mathbb{R}^d}{\mathrm{Argmin}}\,(f).$$

Check this with the (tangent) convexity inequality on S42.

- f strongly convex : f has a unique minimizer.

Check this with the (tangent) strong convexity inequality on S44.

Smooth convex functions

Theorem *The following holds :*

- *If* $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ *is convex, then* $\forall x, z \in \mathbb{R}^d$,

$$\frac{1}{2L}\left\|\nabla f(z) - \nabla f(x)\right\|^2 \leq f(z) - f(x) - \langle \nabla f(x), z - x\rangle \leq \frac{L}{2}\left\|z - x\right\|^2.$$

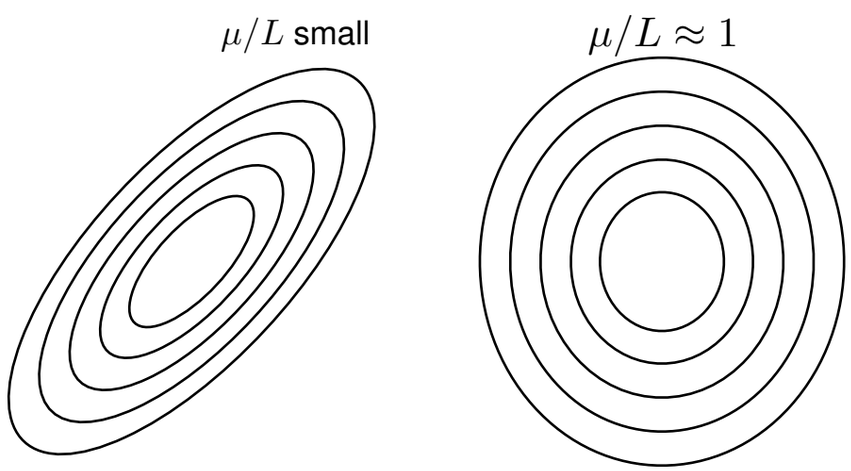
- *If* $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ *is twice differentiable and convex, then* $\forall x \in \mathbb{R}^d$,

$$0 \preceq \nabla^2 f(x) \preceq L \cdot \mathrm{I}.$$

- *If* f *is differentiable and* μ*-strongly convex, then* $\forall x, z \in \mathbb{R}^d$,

$$\frac{\mu}{2}\left\|z - x\right\|^2 \leq f(z) - f(x) - \langle \nabla f(x), z - x\rangle \leq \frac{1}{2\mu}\left\|\nabla f(z) - \nabla f(x)\right\|^2.$$

- *If* $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ *is twice differentiable and* μ*-strongly convex, then* $\forall x \in \mathbb{R}^d$,



$$\mu \cdot \mathrm{I} \preceq \nabla^2 f(x) \preceq L \cdot \mathrm{I}.$$

Smooth convex functions

Proof: We focus on the proof of the first and third claims. The second and and fourth ones follow the same lines as in the proof in S40 , but using the definition of (strong) convexity and Taylor expansion.

Fix $z \in \mathbb{R}^d$ and consider the function $\phi(x) = f(x) - \langle x, \nabla f(z)\rangle$. Obviously, $\phi \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$. It is also convex with z as a global minimum. Thus, applying the descent lemma (see S41) to ϕ, we get

$$\begin{aligned}\phi(z) &\leq \phi\left(x - \frac{1}{L}\nabla\phi(x)\right)\\ &\leq \phi(x) - \frac{1}{L}\langle\nabla\phi(x), x - (x - \nabla\phi(x)/L)\rangle + \frac{L}{2}\left\|x - (x - \nabla\phi(x)/L)\right\|^2 \qquad \leq \phi(x) - \frac{1}{2L}\left\|\nabla\phi(x)\right\|^2.\end{aligned}$$

Thus

$$f(z) - \langle z, \nabla f(z)\rangle \leq f(x) - \langle x, \nabla f(z)\rangle - \frac{1}{2L}\left\|\nabla f(x) - \nabla f(z)\right\|^2$$

which is our claim.

Now, if f is μ-strongly convex, then so is ϕ, and for all $y \in \mathbb{R}^d$,

$$\phi(z) = \min_y \phi(y) \geq \min_y \left\{\phi(x) + \langle\nabla\phi(x), y - x\rangle + \frac{\mu}{2}\left\|y - x\right\|^2\right\}.$$

The minimum on the rhs is attained at which $y - x = -\nabla\phi(x)/\mu$, which leads to

$$\phi(z) \geq \phi(x) - \frac{1}{2\mu}\left\|\nabla\phi(x)\right\|^2.$$

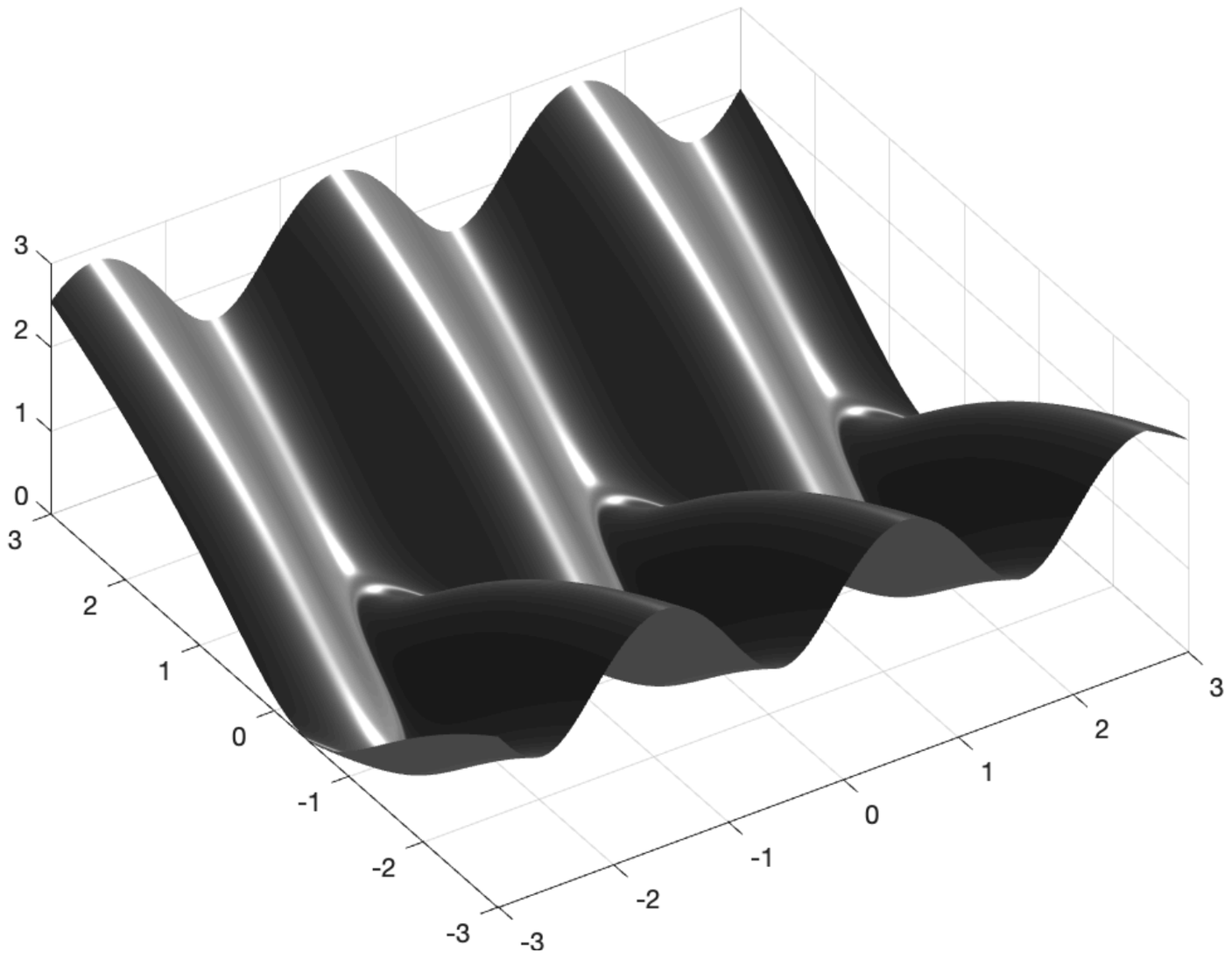
Replacing ϕ and $\nabla\phi$ by their expressions leads the claim. ■

Global quadratic error bound

- As a corollary of the previous theorem, if f is differentiable and μ-strongly convex, then for $x^\star \in \operatorname{Argmin}(f)$ (denote $f^\star = \min f$) :

$$f(x) - f^\star \leq \frac{1}{2\mu} \|\nabla f(x)\|^2, \quad \forall x \in \mathbb{R}^d.$$

- This is known as a global Łojasiewicz property (with Łojasiewicz exponent $1/2$) : we will write $f \in \text{Ł}(1/2)$.
- In fact, it also holds in the non-convex case, e.g. $f(x_1, x_2) = (\sin(x_1) - x_2)^2$.
- In ML, many call it (somehow unduly) the Polyak-Łojasiewicz property.

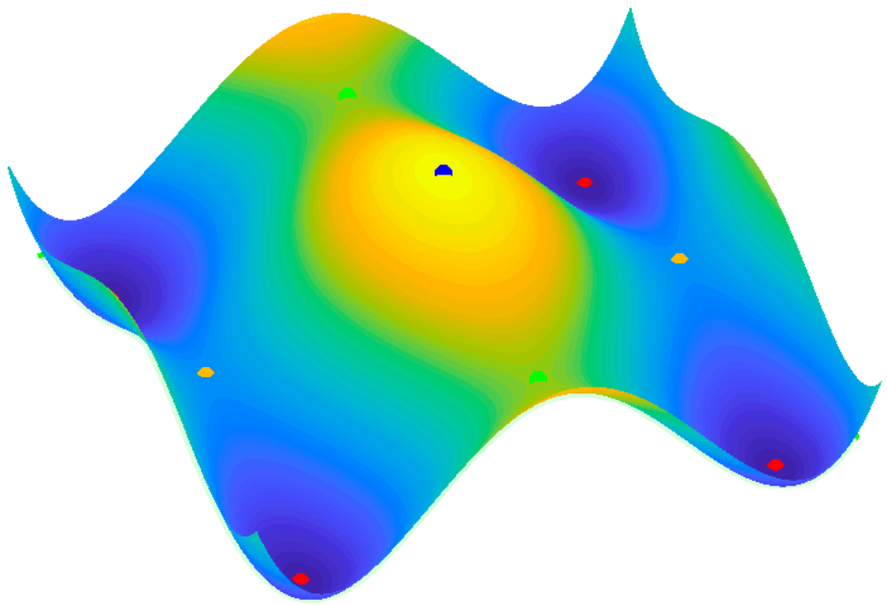


Global quadratic error bound

Definition *A function* $f \in C^2(\mathbb{R}^d)$ *will be Morse if it satisfies the following conditions :*

- *For each critical point* $\hat{x}$, $\nabla^2 f(\hat{x})$ *is nonsingular.*
- *There exists a nonempty set* $I \subseteq \mathbb{N}$ *and* $(\hat{x}_k)_{k\in I}$ *such that* $\mathrm{crit}(f) = \bigcup_{k\in I}\{\hat{x}_k\}$.

- Morse functions are generic in the Baire category sense in the space of C^2 functions.
- Morse functions are Ł$(1/2)$ around each critical point.



Global quadratic error bound

Proposition *If $f \in Ł(1/2)$, then it satisfies the quadratic growth condition (quadratic error bound)*

$$f(x) - f^\star \geq \frac{\mu}{2}\mathrm{dist}(x, \mathrm{Argmin}\ (f))^2$$

for all x close to $\mathrm{Argmin}\ (f)$.

Global quadratic error bound

Proposition *If $f \in Ł(1/2)$, then it satisfies the quadratic growth condition (quadratic error bound)*

$$f(x) - f^\star \geq \frac{\mu}{2}\mathrm{dist}(x, \mathrm{Argmin}\ (f))^2$$

for all x close to $\mathrm{Argmin}\ (f)$.

Proof: For any $y \in \mathbb{R}^d$, consider the solution trajectory to the ODE $\dot{x}(t) = -\nabla f(x(t))$, with initial condition $x(0) = y$, which has a global classical solution by the Cauchy-Lipschitz theorem. Define $\Delta f(x) = f(x) - \min f$. We have for any $t \geq 0$

$$\frac{d}{dt}\sqrt{\Delta f(x(t))} = \frac{\langle \dot{x}(t), \nabla f(x(t))\rangle}{2\sqrt{\Delta f(x(t))}} = -\frac{\|\nabla f(x(t))\|^2}{2\sqrt{\Delta f(x(t))}} \leq -\sqrt{\mu/2}\,\|\nabla f(x(t))\| = -\sqrt{\mu/2}\,\|\dot{x}(t)\| \tag{1}$$

$$\leq -\mu\sqrt{\Delta f(x(t))}, \tag{2}$$

where we used that $f \in Ł(1/2)$ twice. Applying Grönwall inequality to (2) shows that $f(x(t)) \to \min f$ exponentially as $t \to +\infty$. Now integrating (1) from τ to s for any $0 \leq \tau < s$, we get

$$\|x(s) - x(\tau)\| = \left\|\int_\tau^s \dot{x}(t)dt\right\| \leq \int_\tau^s \|\dot{x}(t)\|\,dt \leq -\sqrt{\frac{2}{\mu}}\int_\tau^s \frac{d}{dt}\sqrt{\Delta f(x(t))}dt = \sqrt{\frac{2}{\mu}}\left(\sqrt{\Delta f(x(\tau))} - \sqrt{\Delta f(x(s))}\right) \tag{3}$$

This shows that $\dot{x}(\cdot) \in L^1([0, +\infty[)$ and thus $x(\cdot)$ has the Cauchy property. Hence $x(\cdot)$ converges as $t \to +\infty$ to say $\bar{x}$. The latter is necessarily a global minimizer as we have already shown that $f(x(t)) \to \min f$. Taking $\tau = 0$ and $s \to +\infty$ in (3) entails that

$$\mathrm{dist}(y, \mathrm{Argmin}\ f) \leq \|y - \bar{x}\| \leq \sqrt{\frac{2}{\mu}}(f(y) - \min f).$$

Outline

- Classes of functions.
- Toolbox on sequences.
- Deterministic smooth optimization.
- Stochastic approximation à la Robbins-Monro.
- Stochastic gradient descent: vanishing step-size.
- Stochastic gradient descent for finite sums.

A bit of notations

- We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space with set of events Ω, σ-algebra $\mathcal{F}$, and probability measure $\mathbb{P}$.
- The Borel σ-algebra on $\mathbb{R}^d$ is denoted $\mathcal{B}$.
- A $\mathbb{R}^d$-valued random variable is a measurable map $x : (\Omega, \mathcal{F}) \to (\mathbb{R}^d, \mathcal{B})$.
- A *filtration* $\mathscr{F} = (\mathcal{F}_k)_{k\in\mathbb{N}}$ is a sequence of sub-σ-algebras which satisfies $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all $k \in \mathbb{N}$.
- Given a set of random variables $\{a_0, \ldots, a_k\}$, we denote by $\sigma(a_0, \ldots, a_k)$ the σ-algebra generated by $a_0, \ldots, a_k$. Typically, for a stochastic iterative algorithm, $\mathcal{F}_k = \sigma(a_0, \ldots, a_k)$ is the information up to iteration k.
- An expression (P) is said to hold $(\mathbb{P}\text{-a.s.})$ if $\mathbb{P}(\{\omega \in \Omega : (P) \text{ holds}\}) = 1$.
- Throughout the class, both equalities and inequalities involving random quantities should be understood as holding $\mathbb{P}$-almost surely, whether or not it is explicitly written.

A bit of notations

- We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space with set of events Ω, σ-algebra $\mathcal{F}$, and probability measure $\mathbb{P}$.
- The Borel σ-algebra on $\mathbb{R}^d$ is denoted $\mathcal{B}$.
- A $\mathbb{R}^d$-valued random variable is a measurable map $x : (\Omega, \mathcal{F}) \to (\mathbb{R}^d, \mathcal{B})$.
- A *filtration* $\mathscr{F} = (\mathcal{F}_k)_{k\in\mathbb{N}}$ is a sequence of sub-σ-algebras which satisfies $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all $k \in \mathbb{N}$.
- Given a set of random variables $\{a_0, \ldots, a_k\}$, we denote by $\sigma(a_0, \ldots, a_k)$ the σ-algebra generated by $a_0, \ldots, a_k$. Typically, for a stochastic iterative algorithm, $\mathcal{F}_k = \sigma(a_0, \ldots, a_k)$ is the information up to iteration k.
- An expression (P) is said to hold $(\mathbb{P}\text{-a.s.})$ if $\mathbb{P}(\{\omega \in \Omega : (P) \text{ holds}\}) = 1$.
- Throughout the class, both equalities and inequalities involving random quantities should be understood as holding $\mathbb{P}$-almost surely, whether or not it is explicitly written.

Definition *Given a filtration $\mathscr{F}$, we denote by $\ell_+(\mathscr{F})$ the set of sequences of $[0, +\infty[$-valued random variables $(a_k)_{k\in\mathbb{N}}$ such that, for each $k \in \mathbb{N}$, a_k is $\mathcal{F}_k$ measurable. Then, we also define the following set,*

$$\ell_+^1(\mathscr{F}) \stackrel{\text{def}}{=} \left\{ (a_k)_{k\in\mathbb{N}} \in \ell_+(\mathscr{F}) : \sum_{k\in\mathbb{N}} a_k < +\infty \ (\mathbb{P}\text{-a.s.}) \right\}.$$

Modes of convergence

- Our goal for algorithms : show that $x_k \to \bar{x}$ or $\mathrm{dist}(x_k, \underset{\mathbb{R}^d}{\mathrm{Argmin}}\,(f)) \to 0$ or $f(x_k) - \min f \to 0$.
- What sense to be given when x_k is random ?
- Boils down to studying in what sense a random quantity $\delta_k \in \mathbb{R}$ tends to zero :
  - Convergence almost-surely : $\mathbb{P}(\delta_k \to 0) = \mathbb{P}\left(\{\omega \in \Omega :\ \delta_k(\omega) \to 0\}\right) = 1$.
  - Convergence in probability : $\forall \epsilon > 0, \mathbb{P}(|\delta_k| > \epsilon) \to 0$.
  - Convergence in mean $r \geq 1$: $\mathbb{E}(|\delta_k|^r) \to 0$.
- Relationship between convergences :
  - Almost surely $\Rightarrow$ in probability.
  - In mean $\Rightarrow$ in probability (Markov inequality).
  - In probability (sufficiently fast) $\Rightarrow$ almost surely (Borel-Cantelli lemma).
  - In mean (sufficiently fast) $\Rightarrow$ almost surely (Markov inequality+Borel-Cantelli lemma).
  - Almost surely + domination $\Rightarrow$ in mean (dominated convergence theorem).

Robbins-Siegmund lemma

Lemma (Nonnegative almost supermartingales) *Given a filtration* $\mathscr{F}$ *and the sequences of random variables* $(r_k)_{k\in\mathbb{N}} \in \ell_+(\mathscr{F})$, $(a_k)_{k\in\mathbb{N}} \in \ell_+(\mathscr{F})$, $(\alpha_k)_{k\in\mathbb{N}} \in \ell_+^1(\mathscr{F})$ *and* $(\beta_k)_{k\in\mathbb{N}} \in \ell_+^1(\mathscr{F})$, *satisfying,*

$$\mathbb{E}\left[r_{k+1} \mid \mathcal{F}_k\right] \leq (1+\alpha_k)r_k - a_k + \beta_k \qquad (\mathbb{P}\text{-a.s.})$$

then $(a_k)_{k\in\mathbb{N}} \in \ell_+^1(\mathscr{F})$ *and* $(r_k)_{k\in\mathbb{N}}$ *converges* $(\mathbb{P}$-a.s.$)$ *to a random variable valued in* $[0,+\infty[$.

Lemmas on random sequences

Proof: The complete proof can be found in [Robbins-Siegmund 1985]. We here give a sketch. Let

$$\gamma_k = \prod_{i=0}^{k}(1+\alpha_i)^{-1}, r'_k = \gamma_{k-1} r_k, a'_k = \gamma_k a_k, \text{ and } \beta'_k = \gamma_k \beta_k.$$

All these variables are $\mathcal{F}_k$ measurable. Multiplying the inequality of the lemma by γ_k and using that γ_k is non-negative and decreasing, it follows that $(\mathbb{P}\text{-a.s.})$

$$\mathbb{E}\left[r'_{k+1} \mid \mathcal{F}_k\right] \leq r'_k - a'_k + \beta'_k, \tag{1}$$

and

$$\sum_{k\in\mathbb{N}} \beta'_k \leq \sum_{k\in\mathbb{N}} \beta_k < +\infty. \tag{2}$$

Let

$$s_k = r'_k - \sum_{i=0}^{k-1}(\beta'_i - a'_i)). \tag{3}$$

From (1), we have

$$\mathbb{E}\left[s_{k+1} \mid \mathcal{F}_k\right] = \mathbb{E}\left[r'_{k+1} - \sum_{i=0}^{k}(\beta'_i - a'_i) \mid \leq\right] r'_k - \sum_{i=0}^{k-1}(\beta'_i - a'_i) = s_k,$$

and thus, $(s_k)_{k\in\mathbb{N}}$ is a supermartingale. It then follows from the Doob's martingale convergence theorem that

$$\lim_{k\to\infty} s_k \text{ exists and is finite } (\mathbb{P}\text{-a.s.}) .$$

Hence, by (3) and (2), $\lim_{k\to\infty} r'_k$ exists and is finite $(\mathbb{P}\text{-a.s.})$, and $\sum_{k\in\mathbb{N}} a'_k < +\infty$ $(\mathbb{P}\text{-a.s.})$. Now, observe that $1/\gamma_k = \prod_{i=0}^{k}(1+\alpha_i)$ is convergent since $(\alpha_k)_{k\in\mathbb{N}}$ is summable. It then follows from this and

$$r_k = r'_k/\gamma_{k-1}$$

that $(r_k)_{k\in\mathbb{N}}$ is convergent $(\mathbb{P}\text{-a.s.})$. Similarly, since

$$a_k = a'_k/\gamma_k \leq a'_k \prod_{i=0}^{+\infty}(1+\alpha_i),$$

$\sum_{k\in\mathbb{N}} a_k < +\infty$ $(\mathbb{P}\text{-a.s.})$. ■

Lemmas on random sequences

Lemma *Given a filtration $\mathscr{F}$ and the sequences of random variables $(r_k)_{k\in\mathbb{N}} \in \ell_+(\mathscr{F})$, satisfying*

$$\mathbb{E}\left[r_{k+1} \mid \mathcal{F}_k\right] \leq (1-\alpha_k) r_k + \beta_k \qquad (\mathbb{P}\text{-a.s.}) \ ,$$

where

$$\alpha_k \in [0,1], \quad \beta_k \geq 0, \quad \sum_{k\in\mathbb{N}} \alpha_k = +\infty.$$

(i) If $\sum_{k\in\mathbb{N}} \beta_k < +\infty$, then $r_k \to 0$ $(\mathbb{P}$-a.s.$)$.

(ii) If $\frac{\beta_k}{\alpha_k} \to 0$, then $\mathbb{E}[r_k] \to 0$.

Lemmas on random sequences

Proof: (i) Applying the Robbins-Siegmund lemma in S55 (with $a_k = \alpha_k r_k$), we conclude that r_k converges $(\mathbb{P}\text{-a.s.})$ and $\sum_{k\in\mathbb{N}} \alpha_k r_k < +\infty \quad (\mathbb{P}\text{-a.s.})$. Since $(\alpha_k)_{k\in\mathbb{N}}$ is not summable, we have that (easy to prove by a simple contradiction argument)

$$\liminf_{k\to+\infty} r_k = 0 \qquad (\mathbb{P}\text{-a.s.}) \ .$$

But since the limit exists, we get the claim.

(ii) Taking the total expectation on both sides of the inequality, we have

$$\mathbb{E}[r_{k+1}] \leq (1-\alpha_k)\mathbb{E}[r_k] + \beta_k. \tag{1}$$

To lighten notation, denote $b_k = \mathbb{E}[r_k]$. We get from (1) that b_k obeys

$$b_{k+1} \leq b_k - \alpha_k b_k + \beta_k.$$

Let $\theta \in]0,1[$, and denote the two complementary sets

$$I = \{k : \ \beta_k > \theta\alpha_k b_k\}\,, \qquad I^c = \{k : \ \beta_k \leq \theta\alpha_k b_k\}\,.$$

Two cases are possible :

(a) I is finite. Thus, for k large enough, say $k \geq K$, $k \in I^c$ and hence

$$b_{k+1} \leq b_k - (1-\theta)\alpha_k b_k \leq b_k. \tag{2}$$

Thus b_k is non-negative and decreasing, and so it does converge. On the other hand, summing (2) for k larger than K, we have

$$(1-\theta)\sum_{k\geq K} \alpha_k b_k \leq b_K < +\infty.$$

Recalling that $(\alpha_k)_{k\in\mathbb{N}}$ is not summable, we have that $\liminf_{k\to+\infty} b_k = 0$, but since we have proved that b_k converges, we get $b_k \to 0$.

(b) I is infinite. Then for $k \in I$,

$$u_k \leq \frac{\beta_k}{\theta\alpha_k} \to 0,$$

and we passed to the limit since I is infinite.

Lemmas on random sequences

Lemma *Given a filtration* $\mathscr{F}$ *and a sequence of random variables* $(w_k)_{k\in\mathbb{N}} \in \ell_+(\mathscr{F})$ *and a sequence* $(\alpha_k)_{k\in\mathbb{N}} \in \ell_+$ *such that* $(\alpha_k w_k)_{k\in\mathbb{N}} \in \ell_+^1(\mathscr{F})$ *and* $(\alpha_k)_{k\in\mathbb{N}} \notin \ell^1$. *Then* $\liminf_{k\to\infty} w_k = 0$ $(\mathbb{P}$-*a.s.*$)$. *Assume, moreover, that there exists a constant* $\nu > 0$ *such that*

$$w_k - \mathbb{E}\left[w_{k+1} \mid \mathcal{F}_k\right] \leq \nu\alpha_k \qquad (\mathbb{P}\text{-a.s.})$$

for every $k \in \mathbb{N}$, *then*

$$\lim_k w_k = 0 \qquad (\mathbb{P}\text{-a.s.}) \ .$$

Lemmas on random sequences

Proof: Let $\theta > 0$, and define the complementary sets

$$I = \{k \in \mathbb{N} : \ w_k \leq \theta \ (\mathbb{P}\text{-a.s.})\} \quad I^c = \{k \in \mathbb{N} : \ w_k > \theta \ (\mathbb{P}\text{-a.s.})\}.$$

By our assumptions, we know that there exists a subsequence $\left(w_{k_j}\right)_{j\in\mathbb{N}}$ such that $\lim_{j\to\infty} w_{k_j} = 0 \ (\mathbb{P}\text{-a.s.})$, and thus I is infinite. Since $(\alpha_k w_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathscr{F})$, we also have

$$+\infty > \sum_{k\in\mathbb{N}} \alpha_k w_k \geq \sum_{k\in I^c} \alpha_k w_k \geq \theta \sum_{k\in I^c} \alpha_k \qquad (\mathbb{P}\text{-a.s.}) \ .$$

Thus, for all $\epsilon > 0$, there exists k_ϵ such that

$$\theta \sum_{k\in I^c, k\geq k_\epsilon} \alpha_k \leq \epsilon^2/(4\nu) \qquad (\mathbb{P}\text{-a.s.}) \ .$$

Taking $\theta = \epsilon/2$, this shows that

$$\sum_{k\in I^c, k\geq k_\epsilon} \alpha_k \leq \epsilon/(2\nu) \qquad (\mathbb{P}\text{-a.s.}) \ . \tag{1}$$

Now, for all $k \geq k_\epsilon$, there are two possibilities :

(a) $k \in I$, which is equivalent to $w_k \leq \epsilon/2 < \epsilon \ (\mathbb{P}\text{-a.s.})$.

(b) $k \in I^c$. Let

$$m = \min\{j \in I : \ j \geq k\},$$

which exists since I is infinite. Hence, we have

$$\begin{aligned}
w_k &= (w_k - \mathbb{E}[w_m \mid \mathcal{F}_k]) + \mathbb{E}[w_m \mid \mathcal{F}_k] \\
(w_k \text{ known cond. on } \mathcal{F}_k) \quad &= \mathbb{E}[w_k - w_m \mid \mathcal{F}_k] + \mathbb{E}[w_m \mid \mathcal{F}_k] \\
(\text{telescopicity and conditional expectation}) \quad &= \mathbb{E}\left[\sum_{l=k}^{m-1} w_l - \mathbb{E}[w_{l+1} \mid \mathcal{F}_l] \mid \mathcal{F}_k\right] + \mathbb{E}[w_m \mid \mathcal{F}_k] \\
(\text{by assumption on } w_k) \quad &\leq \nu \sum_{l=k}^{m-1} \alpha_l + \epsilon/2 \\
(I^c \ni k, m-1 \geq k_\epsilon, \text{ and (1)}) \quad &\leq \nu \sum_{l\in I^c, l\geq k_\epsilon} \alpha_l + \epsilon/2 \leq \epsilon \qquad (\mathbb{P}\text{-a.s.}) \ .
\end{aligned}$$

In both cases, we have shown that $(\mathbb{P}\text{-a.s.})$, for all $\epsilon > 0$, there exists k_ϵ such that for all $k \geq k_\epsilon$

$$w_k \leq \epsilon$$

which is nothing but $w_k \to 0 \ (\mathbb{P}\text{-a.s.})$. This concludes the proof. ■

Chung lemmas

Lemma *Let* $r_k \geq 0$ *obeying*

$$r_{k+1} \leq \left(1 - \frac{c}{k}\right) r_k + \frac{c'}{k^{p+1}}, \quad p, c, c' > 0.$$

Then

$$r_k \leq c'(c-p)^{-1} k^{-p} + o\left(k^{-p}\right) \qquad \text{if } c > p,$$

$$r_k = O\left(\frac{\log(k)}{k^p}\right) \qquad \text{if } c = p,$$

$$r_k = O(k^{-c}) \qquad \text{if } p > c.$$

Chung lemmas

Lemma *Let* $r_k \geq 0$ *obeying*

$$r_{k+1} \leq \left(1 - \frac{c}{k}\right) r_k + \frac{c'}{k^{p+1}}, \quad p, c, c' > 0.$$

Then

$$r_k \leq c'(c-p)^{-1} k^{-p} + o\left(k^{-p}\right) \qquad \text{if } c > p,$$

$$r_k = O\left(\frac{\log(k)}{k^p}\right) \qquad \text{if } c = p,$$

$$r_k = O(k^{-c}) \qquad \text{if } p > c.$$

Lemma *Let* $r_k \geq 0$ *obeying*

$$r_{k+1} \leq \left(1 - \frac{c}{k^s}\right) r_k + \frac{c'}{k^t}, \quad s \in]0, 1[, s < t.$$

Then

$$r_k \leq \frac{c'}{c} k^{-(t-s)} + o\left(k^{-(t-s)}\right).$$

Chung lemmas

Lemma *Let* $r_k \geq 0$ *obeying*

$$r_{k+1} \leq \left(1 - \frac{c}{k}\right) r_k + \frac{c'}{k^{p+1}}, \quad p, c, c' > 0.$$

Then

$$r_k \leq c'(c-p)^{-1} k^{-p} + o\left(k^{-p}\right) \qquad \text{if } c > p,$$

$$r_k = O\left(\frac{\log(k)}{k^p}\right) \qquad \text{if } c = p,$$

$$r_k = O(k^{-c}) \qquad \text{if } p > c.$$

Lemma *Let* $r_k \geq 0$ *obeying*

$$r_{k+1} \leq \left(1 - \frac{c}{k^s}\right) r_k + \frac{c'}{k^t}, \quad s \in]0, 1[, s < t.$$

Then

$$r_k \leq \frac{c'}{c} k^{-(t-s)} + o\left(k^{-(t-s)}\right).$$

Proof: See Lemma 4 and Lemma 5 in [Polyak 1985, Chapter 2]. ■

Outline

- Classes of functions.
- Toolbox on sequences.
- Deterministic smooth optimization.
- Stochastic approximation à la Robbins-Monro.
- Stochastic gradient descent: vanishing step-size.
- Stochastic gradient descent for finite sums.

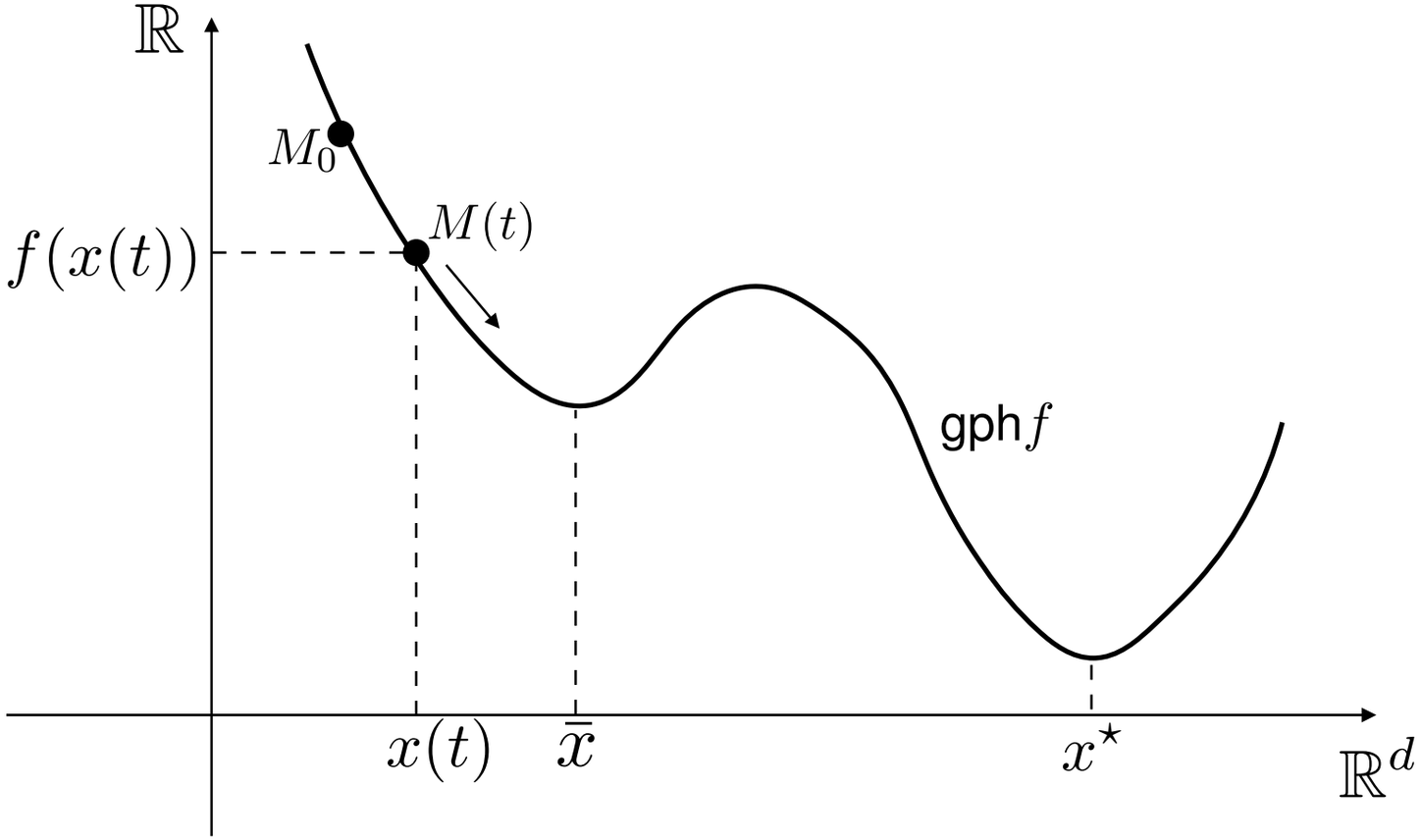
Gradient flow

$$\min_{x\in\mathbb{R}^d} f(x), \quad f\in\mathscr{C}^1(\mathbb{R}^d)$$

- Gradient descent dynamic ([Cauchy 1847]) : $t\in[0,+\infty[$

$$\dot{x}(t)+\nabla f(x(t))=0.$$

  - Velocity = (opposite) of the gradient.
  - Gradient : a force deriving from the potential energy.



Gradient descent

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^1(\mathbb{R}^d)$$

- Gradient descent dynamic ([Cauchy 1847]) : $t \in [0, +\infty[$

$$\dot{x}(t) + \nabla f(x(t)) = 0.$$

- Temporal discretization :

$$\frac{x_{k+1} - x_k}{\gamma_k} = -\nabla f(x_k), \qquad \gamma_k \geq 0$$

Gradient descent

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^1(\mathbb{R}^d)$$

- Gradient descent dynamic ([Cauchy 1847]) : $t \in [0, +\infty[$

$$\dot{x}(t) + \nabla f(x(t)) = 0.$$

- Temporal discretization :

$$\frac{x_{k+1} - x_k}{\gamma_k} = -\nabla f(x_k), \qquad \gamma_k \geq 0$$

Input : gradient function ∇f, step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule ;

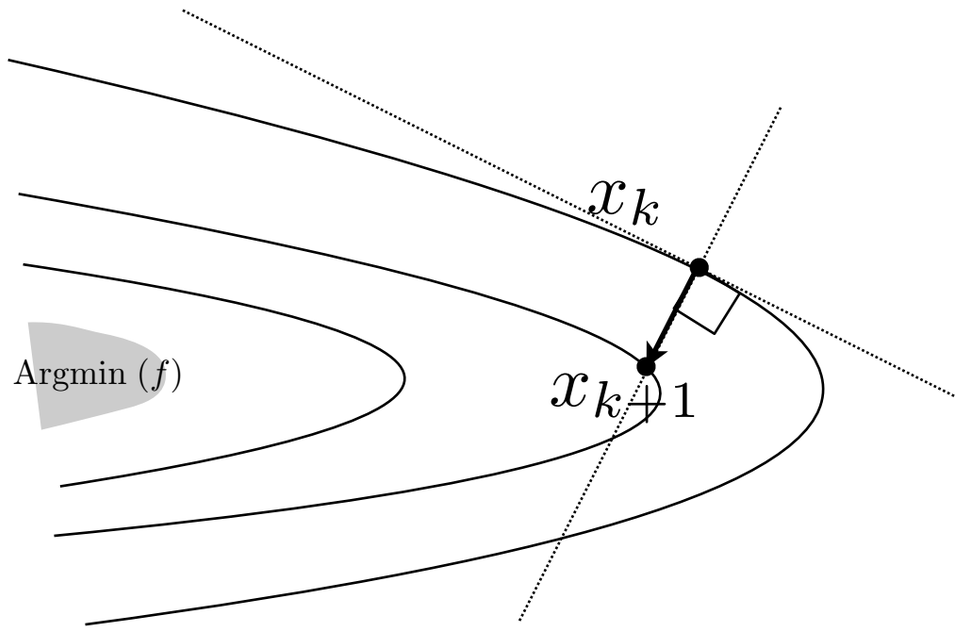
Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  $x_{k+1} = x_k - \gamma_k \nabla f(x_k)$;

  $k \leftarrow k + 1$.

return x_k.



Gradient descent

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^1(\mathbb{R}^d)$$

- Gradient descent dynamic ([Cauchy 1847]) : $t \in [0, +\infty[$

$$\dot{x}(t) + \nabla f(x(t)) = 0.$$

- Temporal discretization :

$$\frac{x_{k+1} - x_k}{\gamma_k} = -\nabla f(x_k), \qquad \gamma_k \geq 0$$

Input : gradient function ∇f, step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule ;

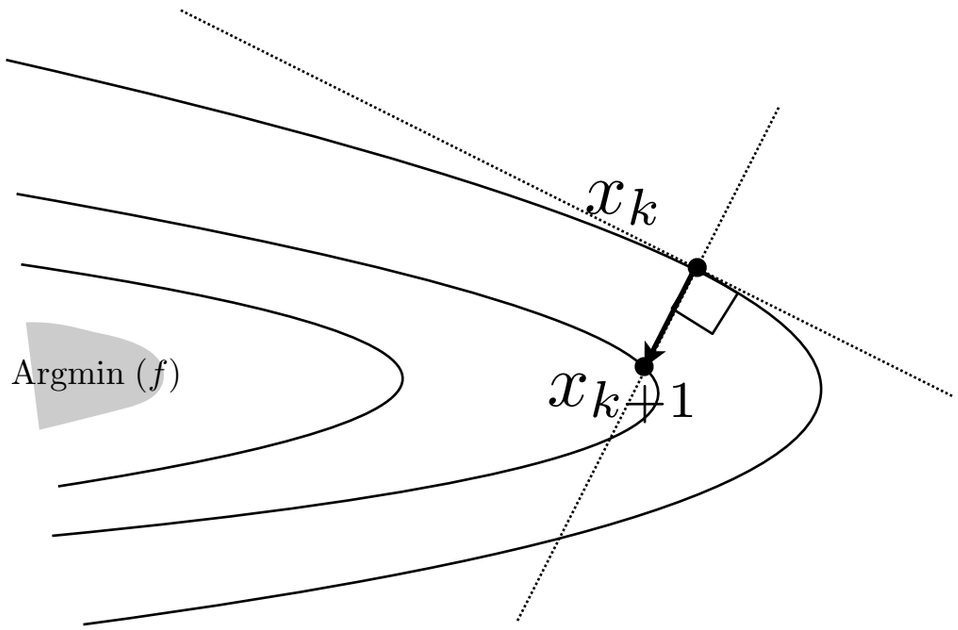
Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  $x_{k+1} = x_k - \gamma_k \nabla f(x_k)$;

  $k \leftarrow k + 1$.

return x_k.



- Simple, yet efficient and most widely used algorithm.
- Its cost/iteration: dominated by the gradient computation.
- In ML with finite sums: n times the gradient of the loss (hence the motivation of stochastic versions).

Gradient descent

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^1(\mathbb{R}^d)$$

Input : gradient function ∇f, step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule ;

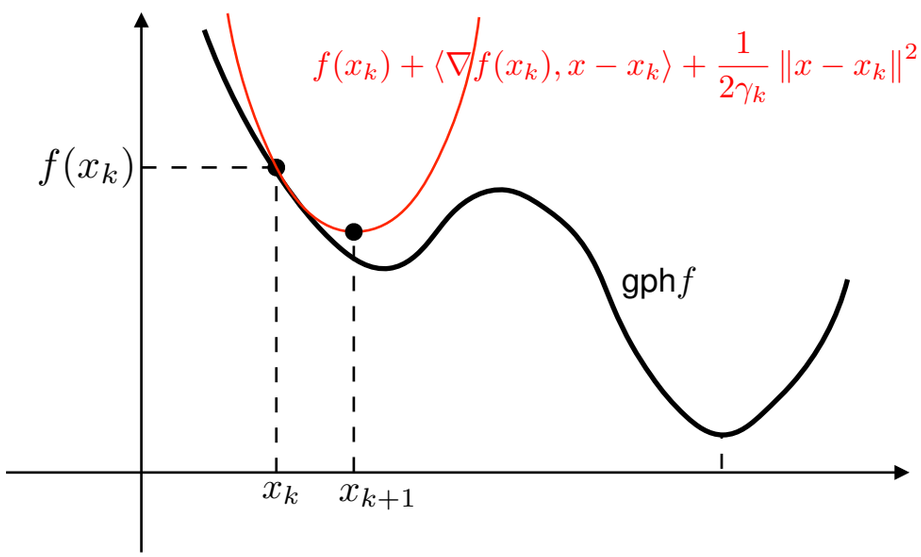
Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

$x_{k+1} = x_k - \gamma_k \nabla f(x_k)$;

$k \leftarrow k+1$.

return x_k.



- Another useful view of gradient descent update :

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathbb{R}^d} f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\gamma_k} \|x - x_k\|^2 .$$

- i.e., approximate f by a quadratic function and then optimize, and repeat.
- For $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ and $\gamma_k \equiv 1/L$, the quadratic approximation is actually a majorant : remember the descent lemma in S41 .

Gradient descent

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^1(\mathbb{R}^d)$$

Input : gradient function ∇f, step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule ;

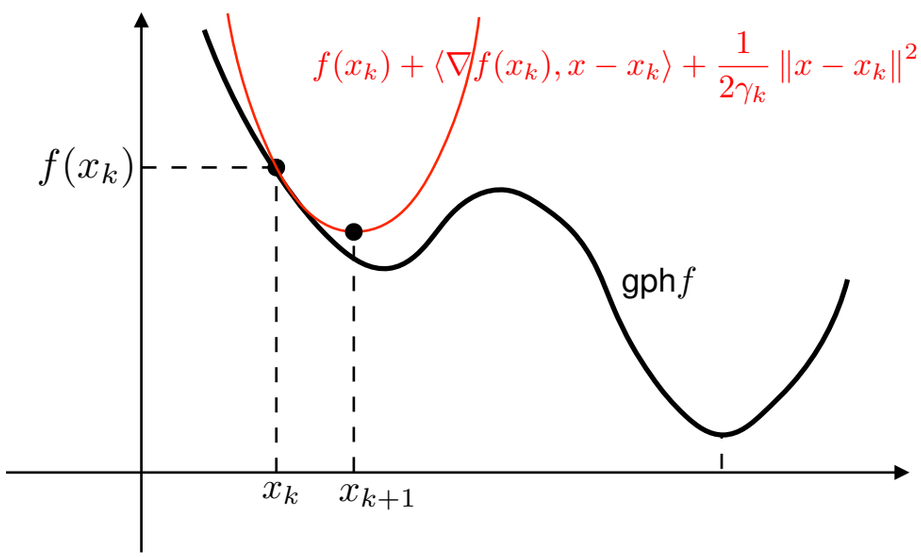
Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

$x_{k+1} = x_k - \gamma_k \nabla f(x_k)$;

$k \leftarrow k+1$.

return x_k.



- Another useful view of gradient descent update :

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathbb{R}^d} f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{2\gamma_k}\|x - x_k\|^2 .$$

- i.e., approximate f by a quadratic function and then optimize, and repeat.
- For $f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d)$ and $\gamma_k \equiv 1/L$, the quadratic approximation is actually a majorant : remember the descent lemma in S41 .
- How to choose the step-size ? Not too large, not too small (e.g. line search, steepest descent or constant step-size).
- When does this algorithm converge ?
  - What quantity does converge (several criteria to characterize convergence) ?
  - At which rate ?
  - Iteration complexity ?

Gradient descent: smooth non-convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$$

Theorem *Suppose that $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ and bounded from below, and that $\gamma_k \equiv \gamma \in]0, 2/L[$. Then*

(i) $f(x_k)$ *converges.*

(ii) $\sum_{k\in\mathbb{N}} \|\nabla f(x_k)\|^2 < +\infty$.

(iii) $\nabla f(x_k) \to 0$ *with the rate*

$$\min_{i\in[k-1]} \|\nabla f(x_i)\| \leq \sqrt{\frac{(f(x_0) - \min f)/(\gamma(1 - L\gamma/2))}{k}}.$$

(iv) If $(x_k)_{k\in\mathbb{N}}$ is bounded, then every accumulation point of $(x_k)_{k\in\mathbb{N}}$ is a critical point of f, i.e. $\operatorname{dist}(x_k, \operatorname{Crit}(f)) \to 0$.

(v) If $\operatorname{Argmin}(f) \neq \emptyset$, $f \in$ Ł$(1/2)$ *and* $\gamma = 1/L$, *then*

S51

$$f(x_k) - \min f \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - \min f) \leq \exp^{-\frac{\mu}{L}k}(f(x_0) - \min f),$$

$$x_k \to x^\star \in \operatorname{Argmin}(f) \text{ at the rate } \|x_k - x^\star\|^2 \leq \exp^{-\frac{\mu}{L}k} \frac{2}{\mu}(f(x_0) - \min f).$$

Gradient descent: smooth non-convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d)$$

Theorem *Suppose that $f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d)$ and bounded from below, and that $\gamma_k \equiv \gamma \in]0, 2/L[$. Then*

(i) $f(x_k)$ *converges.*

(ii) $\sum_{k\in\mathbb{N}} \|\nabla f(x_k)\|^2 < +\infty$.

(iii) $\nabla f(x_k) \to 0$ *with the rate*

$$\min_{i\in[k-1]} \|\nabla f(x_i)\| \leq \sqrt{\frac{(f(x_0) - \min f)/(\gamma(1 - L\gamma/2))}{k}}.$$

(iv) If $(x_k)_{k\in\mathbb{N}}$ is bounded, then every accumulation point of $(x_k)_{k\in\mathbb{N}}$ is a critical point of f, i.e. $\mathrm{dist}(x_k, \mathrm{Crit}(f)) \to 0$.

(v) If $\mathrm{Argmin}\,(f) \neq \emptyset$, $f \in$ Ł$(1/2)$ *and* $\gamma = 1/L$, *then*

S51

$$f(x_k) - \min f \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - \min f) \leq \exp^{-\frac{\mu}{L}k}(f(x_0) - \min f),$$

$$x_k \to x^\star \in \mathrm{Argmin}\,(f) \text{ at the rate } \|x_k - x^\star\|^2 \leq \exp^{-\frac{\mu}{L}k} \frac{2}{\mu}(f(x_0) - \min f).$$

- In general : one needs $k \geq (f(x_0) - \min f)/(\gamma(1 - L\gamma/2))\varepsilon^{-2}$ to achieve precision ε in the gradient $\Rightarrow$ at least $O(\varepsilon^{-2})$ gradient evaluations.
- Under the the 1/2-Łojasiewicz property : one needs $k \gtrsim \frac{L}{\mu}\log(\varepsilon^{-1})$ to achieve precision ε on f and $\mathrm{dist}(\cdot, \mathrm{Argmin}\,(f))^2 \Rightarrow$ at least $O(\log(\varepsilon^{-1}))$ gradient evaluations.
- Coercivity of f is a sufficient condition for boundedness of $(x_k)_{k\in\mathbb{N}}$.

Gradient descent: smooth non-convex

Proof: (i) We have

(Descent lemma in S41)
$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2$$

(Gradient descent)
$$= f(x_k) + \langle \nabla f(x_k), -\gamma \nabla f(x_k) \rangle + \frac{L}{2} \|-\gamma \nabla f(x_k)\|^2$$
$$= f(x_k) - \gamma(1 - L\gamma/2) \|\nabla f(x_k)\|^2 . \tag{1}$$

By the condition on the step-size $(f_k)_{k\in\mathbb{N}}$ is a decreasing sequence, and since f is bounded from below, $(f_k)_{k\in\mathbb{N}}$ converges.

(ii) Summing inequality (1), we have for all $k \in \mathbb{N}$

$$\gamma(1 - L\gamma/2) \sum_{i=0}^{k-1} \|\nabla f(x_i)\|^2 \leq \sum_{i=0}^{k-1} (f(x_i) - f(x_{i+1})) = f(x_0) - f(x_k) \leq f(x_0) - \min f < +\infty. \tag{2}$$

Taking the limit as $k \to +\infty$, we get the claim.

(iii) $\nabla f(x_k) \to 0$ follows the summability result of (ii). Now from (2), we get

$$\gamma(1 - L\gamma/2)k \min_{i\in[k-1]} \|\nabla f(x_i)\|^2 \leq \gamma(1 - L\gamma/2) \sum_{i=0}^{k-1} \|\nabla f(x_i)\|^2 \leq f(x_0) - \min f.$$

(iv) Since $(x_k)_{k\in\mathbb{N}}$ is bounded, it has convergent subsequences. Let $\left(x_{k_j}\right)_{j\in\mathbb{N}}$ be any convergent subsequence, and $\bar{x}$ its accumulation point. Then by continuity of ∇f and claim (iii), we have

$$\nabla f(\bar{x}) = \nabla f\big(\lim_{j\to\infty} x_{k_j}\big) = \lim_{j\to\infty} \nabla f(x_{k_j}) = \lim_{k\to\infty} \nabla f(x_k) = 0,$$

meaning that $\bar{x} \in \mathrm{Crit}(f)$. Now, since $\mathrm{dist}(\cdot, \mathrm{Crit}(f))$ is continuous because $\mathrm{Crit}(f)$ is closed, we obtain

$$\lim_{j\to+\infty} \mathrm{dist}(x_{k_j}, \mathrm{Crit}(f)) = \mathrm{dist}(\lim_{j\to+\infty} x_{k_j}, \mathrm{Crit}(f)) = \mathrm{dist}(\bar{x}, \mathrm{Crit}(f)) = 0.$$

The limit being unique (0) for any convergent subsequence $\left(x_{k_j}\right)_{j\in\mathbb{N}}$ means that the whole sequence $(\mathrm{dist}(x_k, \mathrm{Crit}(f))$ actually converges to 0. ■

Gradient descent: smooth non-convex

Proof: [Continued] (v) In this case (i.e. $\gamma = 1/L$), (1) reads

$$f(x_k) - \min f \leq f(x_{k-1}) - \min f - \frac{1}{2L} \left\|\nabla f(x_{k-1})\right\|^2. \tag{3}$$

When $f \in Ł(1/2)$ (see S51), we get

$$f(x_k) - \min f \leq (1 - \mu/L)(f(x_{k-1}) - \min f) \leq (1 - \mu/L)^k (f(x_0) - \min f) \leq \exp(-\mu/L \cdot k)(f(x_0) - \min f),$$

where the last inequality comes from the fact that $1 - t \leq e^{-t}$. For the last bound, we invoked the quadratic growth bound in S51

For the convergence of the sequence $(x_k)_{k\in\mathbb{N}}$, we will show that its has a finite length and is thus a Cauchy sequence. Denote for short $\Delta_k \overset{\text{def}}{=} f(x_k) - \min f$. If $\Delta_k = 0$ $x_k \in \operatorname{Argmin}(f)$ for some $k \geq 0$, then this holds for all $i \geq k$, and thus there is nothing to prove. We thus suppose that $\Delta_k \neq 0$. By convexity of $-\sqrt{\cdot}$, we have

$$\begin{aligned}
-\sqrt{\Delta_k} &\geq -\sqrt{\Delta_{k-1}} - \frac{\Delta_k - \Delta_{k-1}}{2\sqrt{\Delta_{k-1}}} \\
&= -\sqrt{\Delta_{k-1}} + \frac{f(x_{k-1}) - f(x_k)}{2\sqrt{\Delta_{k-1}}} \\
\text{(By (3))} \qquad &\geq -\sqrt{\Delta_{k-1}} + \frac{\left\|\nabla f(x_{k-1})\right\|^2}{4L\sqrt{\Delta_{k-1}}} \\
\text{(The } Ł(1/2) \text{ condition in S51)} \qquad &\geq -\sqrt{\Delta_{k-1}} + \frac{\left\|\nabla f(x_{k-1})\right\|^2}{4L\sqrt{\mu/2}\left\|\nabla f(x_{k-1})\right\|} \\
&= -\sqrt{\Delta_{k-1}} + \frac{\sqrt{2/\mu}}{4L}\left\|\nabla f(x_{k-1})\right\| \\
\text{(Gradient descent)} \qquad &= -\sqrt{\Delta_{k-1}} + \sqrt{\frac{1}{8\mu}}\left\|x_k - x_{k-1}\right\|.
\end{aligned}$$

By the telescopic sequence, we have

$$\sum_{k\geq 1} \|x_k - x_{k-1}\| \leq \sqrt{8\mu}\sqrt{\Delta_0} = \sqrt{8\mu}\sqrt{f(x_0) - \min f} < +\infty.$$

This entails that x_k converges to say $\bar{x}$. But we know that accumulation points of $(x_k)_{k\in\mathbb{N}}$ are in $\operatorname{Argmin}(f)$. Indeed, we have by continuity of f that

$$f(\bar{x}) = f(\lim_{k\to+\infty} x_k) = \lim_{k\to+\infty} f(x_k) = \min f \Rightarrow \bar{x} \in \operatorname{Argmin}(f).$$

Since $f \in Ł(1/2)$, this together with S51 implies that $f(x_k) - \min f \geq \mu \left\|x_k - x^\star\right\|^2/2$, which concludes the proof.

Gradient descent: smooth convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d) \text{ and convex}$$

Theorem *Suppose that* $f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d)$*, bounded from below and convex, that* $\operatorname{Argmin}(f) \neq \emptyset$ *and* $\gamma_k \equiv \gamma \in]0, 2/L[$.

Then

(i) $\sum_{k\in\mathbb{N}} k \|\nabla f(x_k)\|^2 < +\infty \Rightarrow \|\nabla f(x_k)\| = o(k^{-1/2})$ *and* $\min_{i\in[k]} \|\nabla f(x_k)\| = O(k^{-1})$.

(ii) $f(x_k)$ *converges to* $\min f$ *at the rate*

$$f(x_k) - \min f = O(1/k).$$

(iii) The sequence of iterates $(x_k)_{k\in\mathbb{N}}$ *converges to a point in* $\operatorname{Argmin}(f)$.

(iv) If $\gamma \in]0, 1/L]$*, then*

$$f(x_k) - \min f \leq \frac{(L/2)\operatorname{dist}(x_0, \operatorname{Argmin}(f))^2}{k} \quad \textit{and} \quad f(x_k) - \min f = o(1/k).$$

Gradient descent: smooth convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d) \text{ and convex}$$

Theorem *Suppose that* $f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d)$*, bounded from below and convex, that* $\mathrm{Argmin}\,(f) \neq \emptyset$ *and* $\gamma_k \equiv \gamma \in]0, 2/L[$*.*

Then

(i) $\sum_{k\in\mathbb{N}} k \left\|\nabla f(x_k)\right\|^2 < +\infty \Rightarrow \left\|\nabla f(x_k)\right\| = o(k^{-1/2})$ *and* $\min_{i\in[k]} \left\|\nabla f(x_k)\right\| = O(k^{-1})$*.*

(ii) $f(x_k)$ *converges to* $\min f$ *at the rate*

$$f(x_k) - \min f = O(1/k).$$

(iii) The sequence of iterates $(x_k)_{k\in\mathbb{N}}$ *converges to a point in* $\mathrm{Argmin}\,(f)$*.*

(iv) If $\gamma \in]0, 1/L]$*, then*

$$f(x_k) - \min f \leq \frac{(L/2)\mathrm{dist}(x_0, \mathrm{Argmin}\,(f))^2}{k} \quad \textit{and} \quad f(x_k) - \min f = o(1/k).$$

- One needs $k \geq C\varepsilon^{-1}$ for some constant $C > 0$ to achieve precision ε on the function values $f \Rightarrow$ at least $O(\varepsilon^{-1})$ gradient evaluations.
- The term $\mathrm{dist}(x_0, \mathrm{Argmin}\,(f))$ may hide dependence on the dimension d.

Gradient descent: smooth convex

Proof: There are several proofs many of which require $\gamma \in]0, 1/L]$. We here provide a general, yet simple, one which is based a Lyapunov analysis and holds for $\gamma \in]0, 2/L[$. To lighten notation, denote $\rho \stackrel{\text{def}}{=} \gamma(1 - \gamma L/2)$. Take any $x^\star \in \operatorname{Argmin}(f)$. Define the sequence :

$$V_k \stackrel{\text{def}}{=} k(f(x_k) - \min f) + \frac{1}{2\gamma} \|x_k - x^\star\|^2 .$$

This is a non-negative sequence. We will now show that it is decreasing. We have

$$V_{k+1} - V_k = k(f(x_{k+1}) - f(x_k)) + f(x_{k+1}) - \min f + \frac{1}{2\gamma}\left(\|x_k - \gamma\nabla f(x_k) - x^\star\|^2 - \|x_k - x^\star\|^2\right)$$

$$= k(f(x_{k+1}) - f(x_k)) + f(x_{k+1}) - \min f + \frac{1}{2\gamma}\left(-2\gamma\langle\nabla f(x_k), x_k - x^\star\rangle + \gamma^2 \|\nabla f(x_k)\|^2\right)$$

(Descent lemma in S41)

$$\leq k\left(\langle\nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2\right) + f(x_{k+1}) - \min f - \langle\nabla f(x_k), x_k - x^\star\rangle + \frac{\gamma}{2}\|\nabla f(x_k)\|^2$$

(Gradient descent step)

$$= -k\rho\|\nabla f(x_k)\|^2 + (f(x_{k+1}) - \min f - \langle\nabla f(x_k), x_k - x^\star\rangle) + \frac{\gamma}{2}\|\nabla f(x_k)\|^2$$

((1) in S67)

$$\leq -k\rho\|\nabla f(x_k)\|^2 + (f(x_k) - \min f - \langle\nabla f(x_k), x_k - x^\star\rangle) - \rho\|\nabla f(x)\|^2 + \frac{\gamma}{2}\|\nabla f(x_k)\|^2$$

(Theorem in S49)

$$\leq -k\rho\|\nabla f(x_k)\|^2 - \frac{1}{2L}\|\nabla f(x_k)\|^2 - \rho\|\nabla f(x_k)\|^2 + \frac{\gamma}{2}\|\nabla f(x_k)\|^2 .$$

Let k_0 such that $k_0 \geq \frac{\gamma L - 1}{\gamma L(2 - \gamma L)} - 1$. Thus

$$V_{k+1} - V_k \leq -(k - k_0)\rho\|\nabla f(x_k)\|^2 - \left((k_0 + 1)\rho + \frac{1}{2L} - \frac{\gamma}{2}\right)\|\nabla f(x_k)\|^2 .$$

Under the assumption on k_0, we have $\left((k_0 + 1)\rho + \frac{1}{2L} - \frac{\gamma}{2}\right) \leq 0$, and thus

$$V_{k+1} - V_k \leq -(k - k_0)\|\nabla f(x_k)\|^2 . \tag{1}$$

Gradient descent: smooth convex

Proof: [Continued]

(i) (1) tells us that $(V_k)_{k\geq k_0}$ is decreasing, and since it it non-negative it converges (and in particular is bounded). Summing (1) over $k \geq k_0$, we have

$$\rho \sum_{k\geq k_0} k \left\|\nabla f(x_k)\right\|^2 \leq V_{k_0} < +\infty.$$

(ii) We have from (1) and for $k \geq k_0$

$$k(f(x_k) - \min f) \leq V_k \leq V_{k_0},$$

from which we get the rate and thus $f(x_k) - \min f \to 0$ by passing to the limit.

(iii) We have by gradient descent and convexity of f

$$\begin{aligned}
\left\|x_{k+1} - x^\star\right\|^2 &= \left\|x_k - x^\star\right\|^2 - 2\gamma\langle\nabla f(x_k), x_k - x^\star\rangle + \gamma^2 \left\|\nabla f(x_k)\right\|^2 \\
\text{(1st item in Theorem in S49)} \quad &\leq \left\|x_k - x^\star\right\|^2 - 2\gamma(f(x_k) - \min f) - \frac{\gamma}{L}\left\|\nabla f(x_k)\right\|^2 + \gamma^2 \left\|\nabla f(x_k)\right\|^2 \\
&= \left\|x_k - x^\star\right\|^2 - 2\gamma(f(x_k) - \min f) - \gamma/L\,(1-\gamma L)\left\|\nabla f(x_k)\right\|^2 \qquad (3) \\
&\leq \left\|x_k - x^\star\right\|^2 - \gamma/L\,(1-\gamma L)\left\|\nabla f(x_k)\right\|^2.
\end{aligned}$$

Let $\rho = \gamma/L\,(1-\gamma L)$. ρ is positive for $\gamma \in]0, 1/L]$ and negative for $\gamma \in]1/L, 2/L[$. Thus

$$\left\|x_{k+1} - x^\star\right\|^2 \leq \left\|x_k - x^\star\right\|^2 + |\rho| \left\|\nabla f(x_k)\right\|^2.$$

We have shown in (i) that $\left(k\left\|\nabla f(x_k)\right\|^2\right)_{k\in\mathbb{N}}$ is summable, and thus so is $\left(\left\|\nabla f(x_k)\right\|^2\right)_{k\in\mathbb{N}}$. Therefore

$$\left\|x_{k+1} - x^\star\right\|^2 \leq \left\|x_k - x^\star\right\|^2 + \beta_k$$

where β_k is summable. It follows from the lemma in S55 that $(\|x_k - x^\star\|)_{k\in\mathbb{N}}$ converges. In particular $(x_k)_{k\in\mathbb{N}}$ is a bounded sequence, and we can then extract converging subsequences. Arguing as in the proof of the non-convex case (see claim (iv) on S66), we easily see that for any convergent subsequence $\left(x_{k_j}\right)_{j\in\mathbb{N}}$, $x_{k_j} \to \bar{x} \in \mathrm{Crit(f)} = \mathrm{Argmin}\ (f)$ (the last identity follows by convexity of f). Thus, since $(\|x_k - x^\star\|)_{k\in\mathbb{N}}$ converges for any $x^\star \in \mathrm{Argmin}\ (f)$, we apply this at $x^\star = \bar{x}$ to infer that

$$0 = \left\|\lim_{j\to\infty} x_{k_j} - \bar{x}\right\| = \lim_{j\to\infty} \left\|x_{k_j} - \bar{x}\right\| = \lim_{k\to\infty} \left\|x_k - \bar{x}\right\|.$$

Gradient descent: smooth convex

Proof: [Continued]

(iv) We have already shown in claim (i) above that $(V_k)_{k\geq k_0}$ converges, and from (iii) that $(\|x_k - x^\star\|)_{k\in\mathbb{N}}$ also converges. It then follows from the definition of V_k that $\lim_{k\to\infty} k(f(x_k) - \min f)$ exists.

We embark from (3) to get

$$\|x_{k+1} - x^\star\|^2 \leq \|x_k - x^\star\|^2 - 2\gamma(f(x_k) - \min f) - \gamma\,(1/L - \gamma)\,\|\nabla f(x_k)\|^2\,.$$

Discarding the last negative term and using the telescopic property, we deduce that

$$\sum_{k\in\mathbb{N}} (f(x_k) - \min f) \leq \|x_0 - x^\star\|^2 < +\infty.$$

Denote $\delta_k \stackrel{\text{def}}{=} k(f(x_k) - \min f)$. We then have

$$\sum_{k\in\mathbb{N}} \frac{\delta_k}{k} < +\infty.$$

Since $(1/k)_{k\in\mathbb{N}}$ is not summable, it follows that $\liminf_{k\to+\infty} \delta_k = 0$. But we have started by precisely showing that δ_k has a limit and thus this limit is 0. In turn,

$$\lim_{k\to+\infty} k(f(x_k) - \min f) = \lim_{k\to+\infty} \delta_k = 0.$$

This completes the proof. ■

Gradient descent: smooth strongly convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \text{ and convex}$$

Theorem *Suppose that* $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ *and strongly convex and* $\gamma_k \equiv \gamma \in 1/L$. *Then*

$$\frac{\mu}{2}\left\|x_k - x^\star\right\|^2 \leq f(x_k) - \min f \leq (1-\mu/L)^k (f(x_0) - \min f).$$

Gradient descent: smooth strongly convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \text{ and convex}$$

Theorem *Suppose that* $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ *and strongly convex and* $\gamma_k \equiv \gamma \in 1/L$. *Then*

$$\frac{\mu}{2}\left\|x_k - x^\star\right\|^2 \leq f(x_k) - \min f \leq (1-\mu/L)^k (f(x_0) - \min f).$$

- One needs $k \geq C\sqrt{L/\mu)}\log(1/\varepsilon)$ for some constant $C > 0$ to achieve precision ε.
- This linear rate can be slightly improved to $(1-\mu/L)/(1+\mu/L)$, We omit the details here.

Gradient descent: smooth strongly convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \text{ and convex}$$

Theorem *Suppose that $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ and strongly convex and $\gamma_k \equiv \gamma \in 1/L$. Then*

$$\frac{\mu}{2}\left\|x_k - x^\star\right\|^2 \leq f(x_k) - \min f \leq (1-\mu/L)^k (f(x_0) - \min f).$$

- One needs $k \geq C\sqrt{L/\mu)}\log(1/\varepsilon)$ for some constant $C > 0$ to achieve precision ε.
- This linear rate can be slightly improved to $(1-\mu/L)/(1+\mu/L)$, We omit the details here.

Proof: Since f is μ-strongly convex it has a unique minimizer $x^\star$ and verifies the 1/2–Łojasiewicz property (see S51). We get the claim from Theorem S66(v). ■

Optimal convergence rate: smooth convex

- First-order method : any iterative algorithm that selects x_k in $x_0 + \mathrm{span}(\nabla f(x_0), \dots, \nabla f(x_{k-1}))$.
- Problem class : convex $\mathscr{C}_L^{1,1}(\mathbb{R}^d)$ functions with a global minimizer $x^\star$.

Theorem **([Nemirovski and Yudin 1983])** *For any $k \leq (d-1)/2$ and any $x_0 \in \mathbb{R}^d$, there exists a convex function $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ such that any first-order algorithm satisfies*

$$f(x_k) - \min f \geq \frac{3L\mathrm{dist}(x_0, \mathrm{Argmin}\ (f))^2}{32(k+1)^2}.$$

Optimal convergence rate: smooth convex

- First-order method : any iterative algorithm that selects x_k in $x_0 + \mathrm{span}(\nabla f(x_0), \dots, \nabla f(x_{k-1}))$.
- Problem class : convex $\mathscr{C}_L^{1,1}(\mathbb{R}^d)$ functions with a global minimizer $x^\star$.

Theorem (**[Nemirovski and Yudin 1983]**) *For any $k \leq (d-1)/2$ and any $x_0 \in \mathbb{R}^d$, there exists a convex function $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ such that any first-order algorithm satisfies*

$$f(x_k) - \min f \geq \frac{3L\mathrm{dist}(x_0, \mathrm{Argmin}\ (f))^2}{32(k+1)^2}.$$

Conclusions

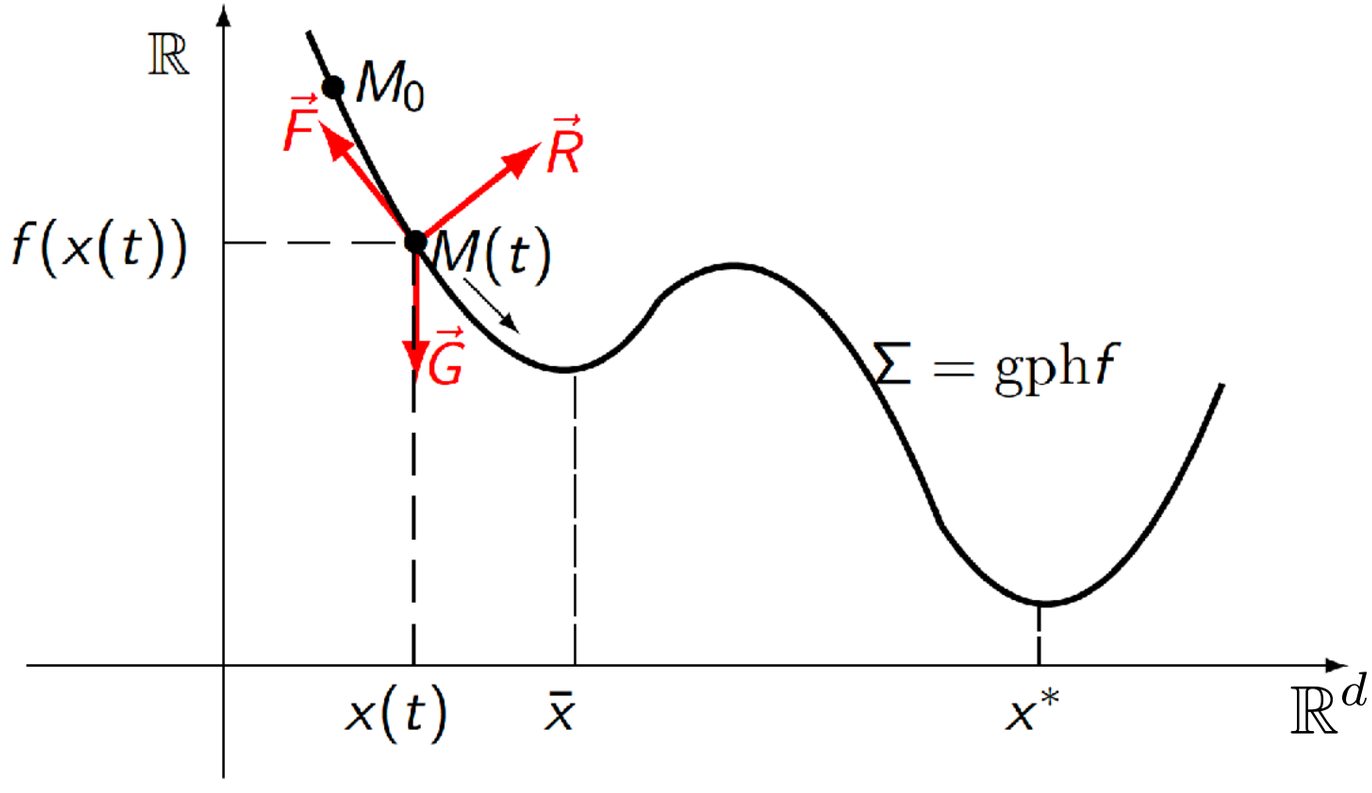
- Gradient descent (rate $O(1/k)$) is not optimal on this class of functions.
- The rate $O(1/k^2)$ is.
- Can we design an algorithm to do so ?
- **Yes**: the key is **inertia** (mostly called momentum in machine learning).

Inertial gradient dynamic

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^1(\mathbb{R}^d)$$

- Inertial dynamic with asymptotically vanishing viscous damping $t \in [t_0, +\infty[$, $t_0 > 0$:

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0, \quad \alpha > 0.$$



Mechanical interpretation : $\vec{F}$: friction. $\vec{R}$: reaction. $\vec{G}$: gravity.

Inertial gradient algorithm

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^1(\mathbb{R}^d)$$

- Inertial dynamic with asymptotically vanishing viscous damping $t \in [t_0, +\infty[$, $t_0 > 0$:

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0, \quad \alpha > 0.$$

- Temporal discretization with time-step $\sqrt{\gamma}$:

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{\gamma} + \frac{\alpha}{k\gamma}(x_k - x_{k-1}) + \nabla f(y_k) = 0, \quad \alpha > 0.$$

Inertial gradient algorithm

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^1(\mathbb{R}^d)$$

- Inertial dynamic with asymptotically vanishing viscous damping $t \in [t_0, +\infty[$, $t_0 > 0$:

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0, \quad \alpha > 0.$$

- Temporal discretization with time-step $\sqrt{\gamma}$:

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{\gamma} + \frac{\alpha}{k\gamma}(x_k - x_{k-1}) + \nabla f(y_k) = 0, \quad \alpha > 0.$$

[Nesterov 1983, 2004]

Input : gradient function ∇f, step-size γ, x_0, x_{-1}, stopping rule ;

Initialization : $k = 0$;

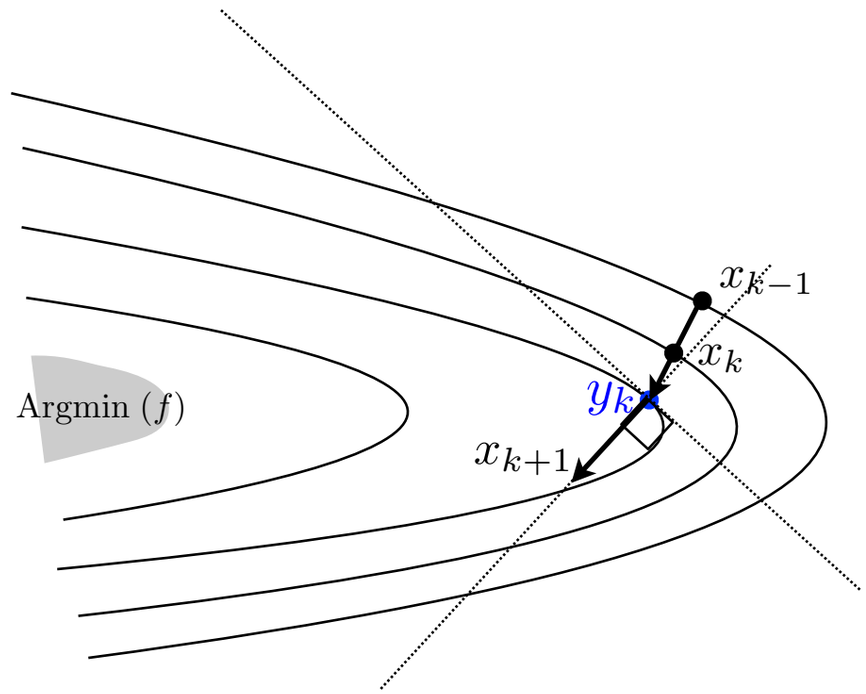
while *Stopping rule not satisfied* **do**

$y_k = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1})$;

$x_{k+1} = y_k - \gamma\nabla f(y_k)$;

$k \leftarrow k + 1$.

return x_k.



Inertial gradient algorithm

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^1(\mathbb{R}^d)$$

- Inertial dynamic with asymptotically vanishing viscous damping $t \in [t_0, +\infty[$, $t_0 > 0$:

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0, \quad \alpha > 0.$$

- Temporal discretization with time-step $\sqrt{\gamma}$:

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{\gamma} + \frac{\alpha}{k\gamma}(x_k - x_{k-1}) + \nabla f(y_k) = 0, \quad \alpha > 0.$$

[Nesterov 1983, 2004]

Input : gradient function ∇f, step-size γ, x_0, x_{-1}, stopping rule ;

Initialization : $k = 0$;

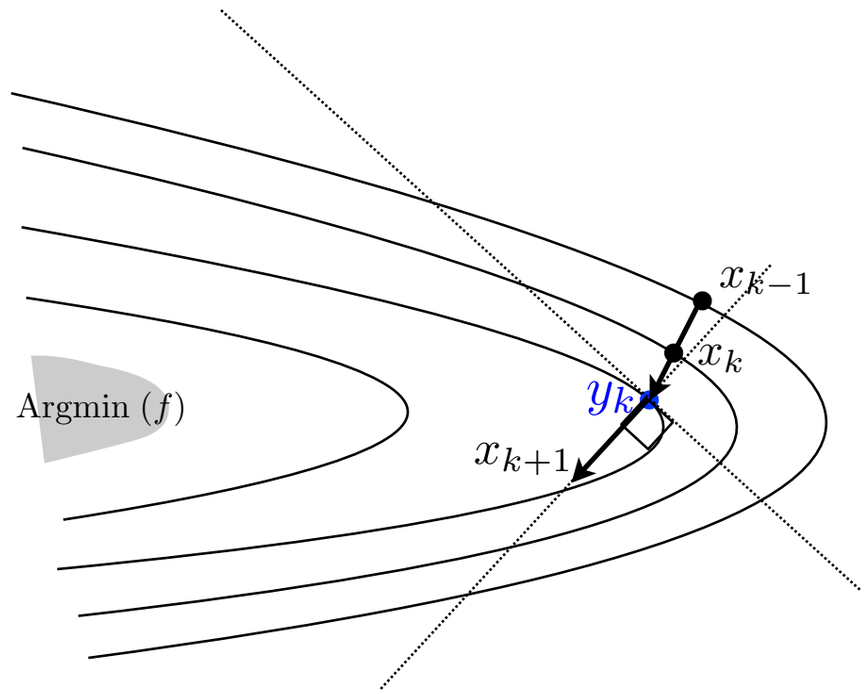
while *Stopping rule not satisfied* **do**

$y_k = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1})$;

$x_{k+1} = y_k - \gamma\nabla f(y_k)$;

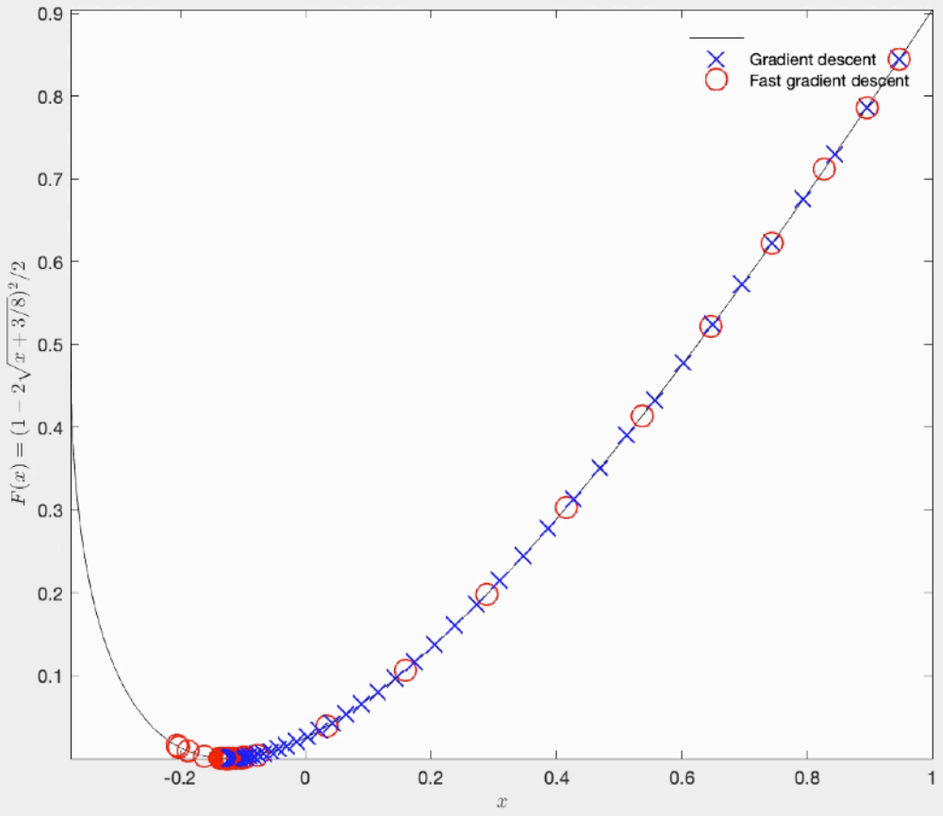
$k \leftarrow k + 1$.

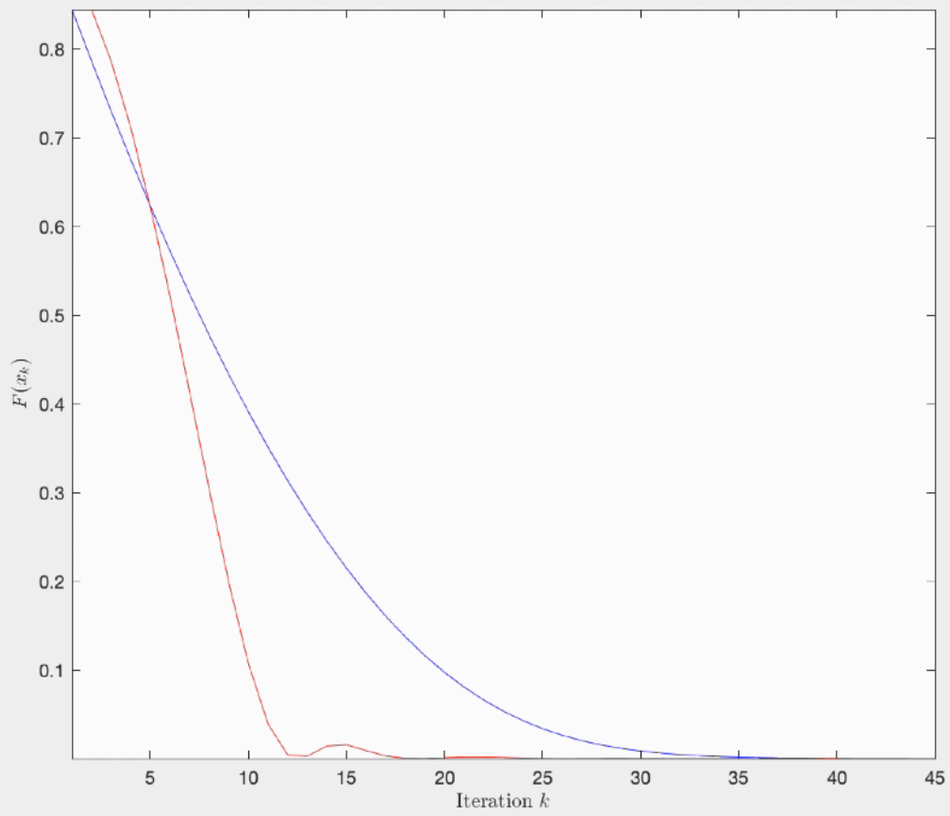
return x_k.



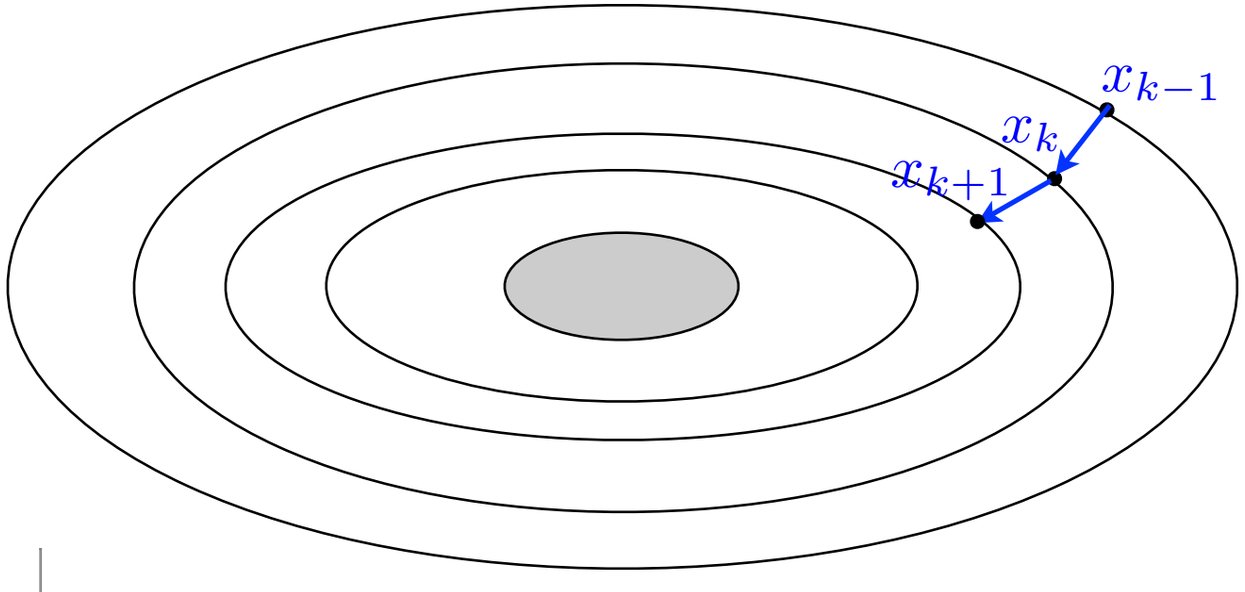
- Same complexity as gradient descent.
- Keep two previous iterates.
- Its cost/iteration: dominated by the gradient computation.
- In ML with finite sums: n times the gradient of the loss (hence the motivation of stochastic versions).

Gradient descent vs Inertial version

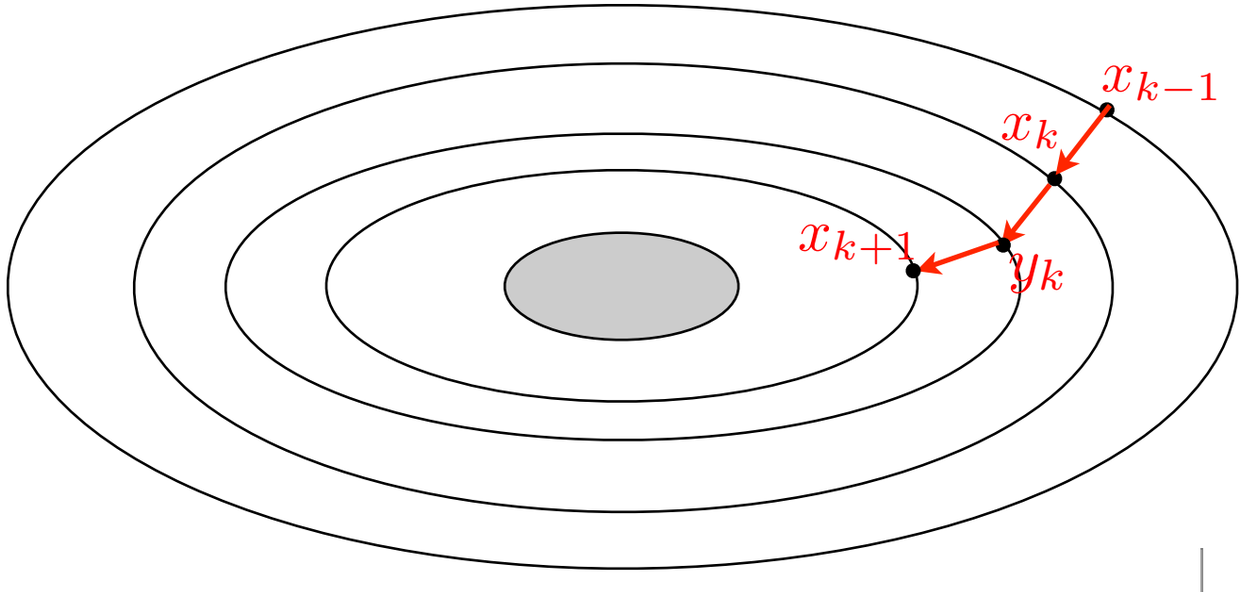




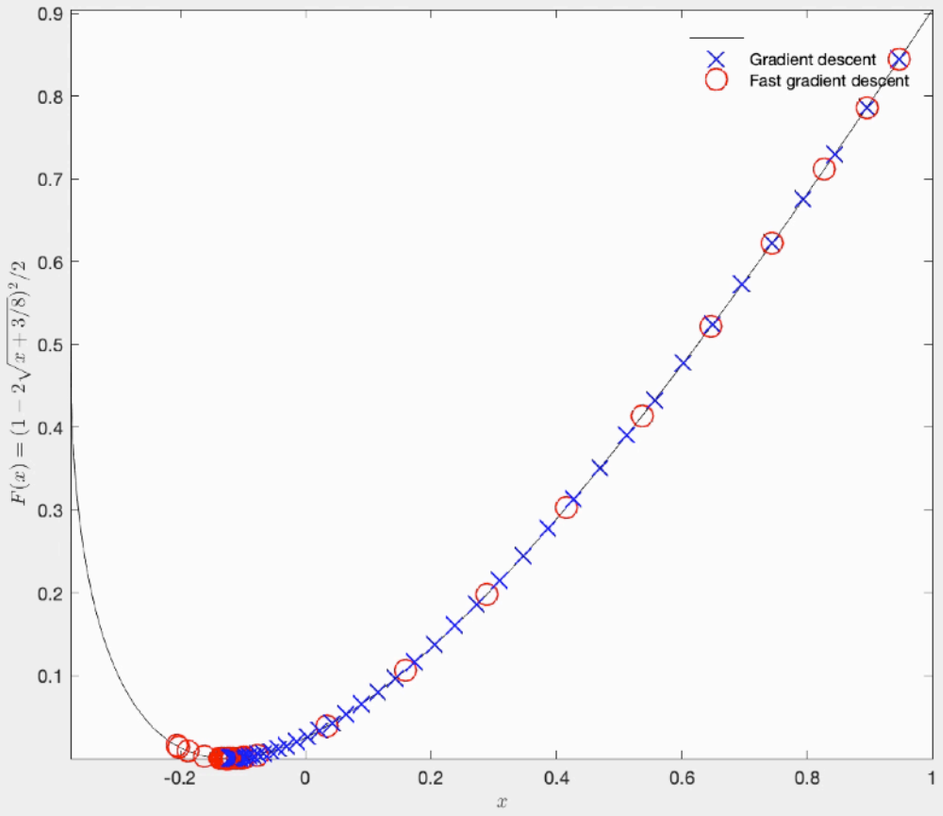
Gradient descent

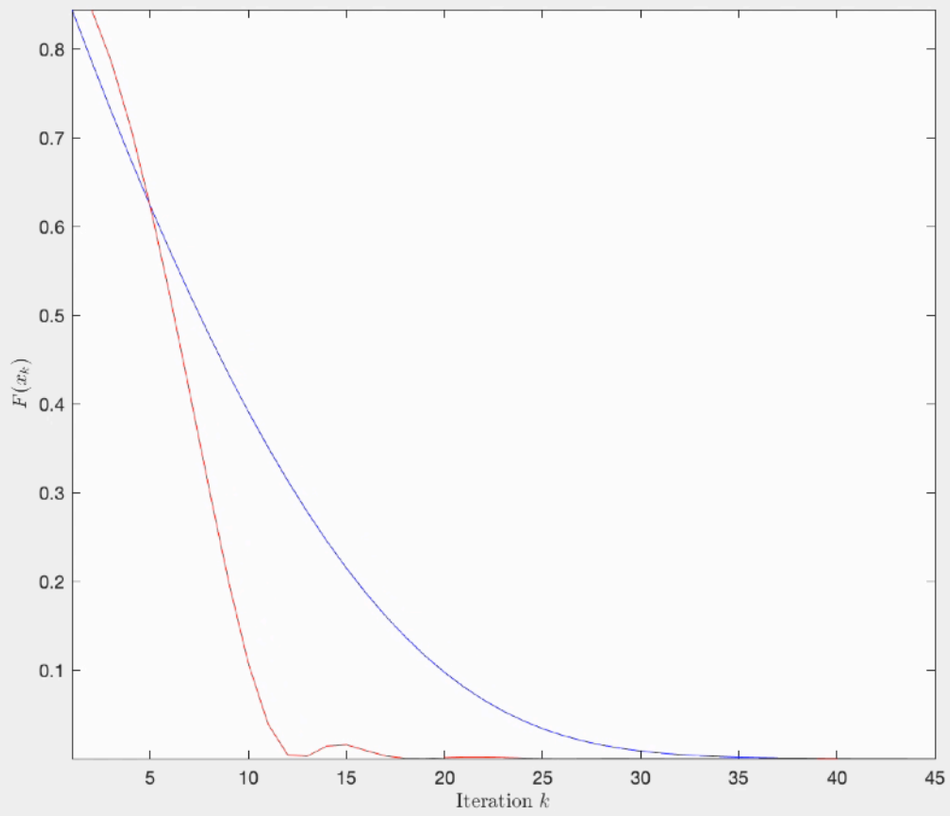


Inertial gradient descent

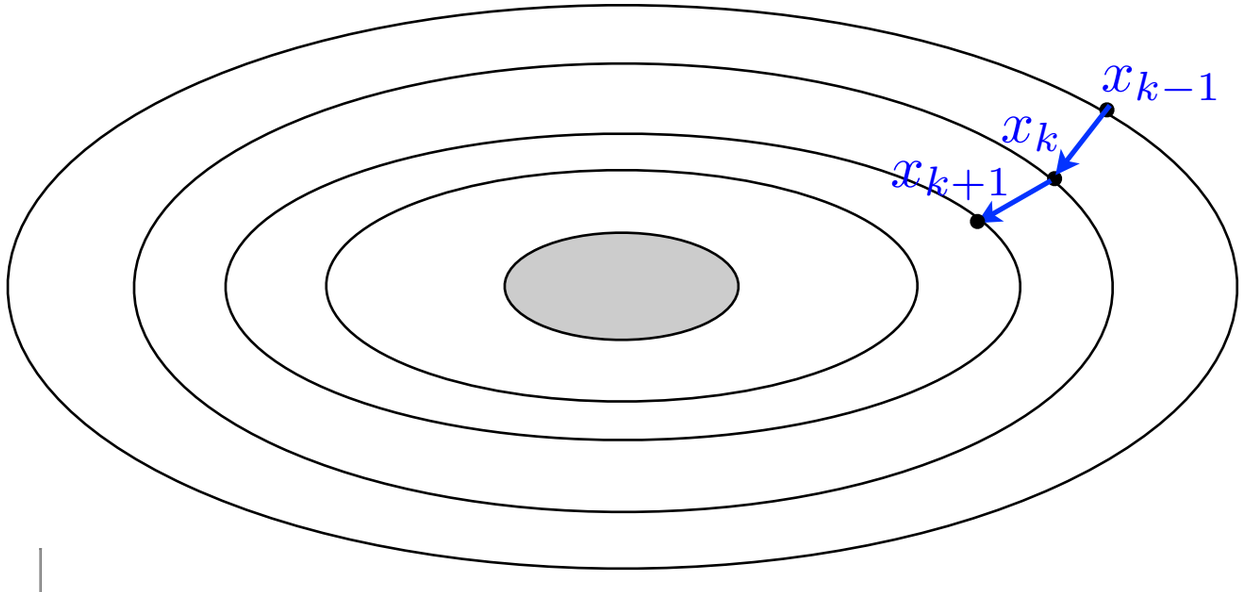


Gradient descent vs Inertial version

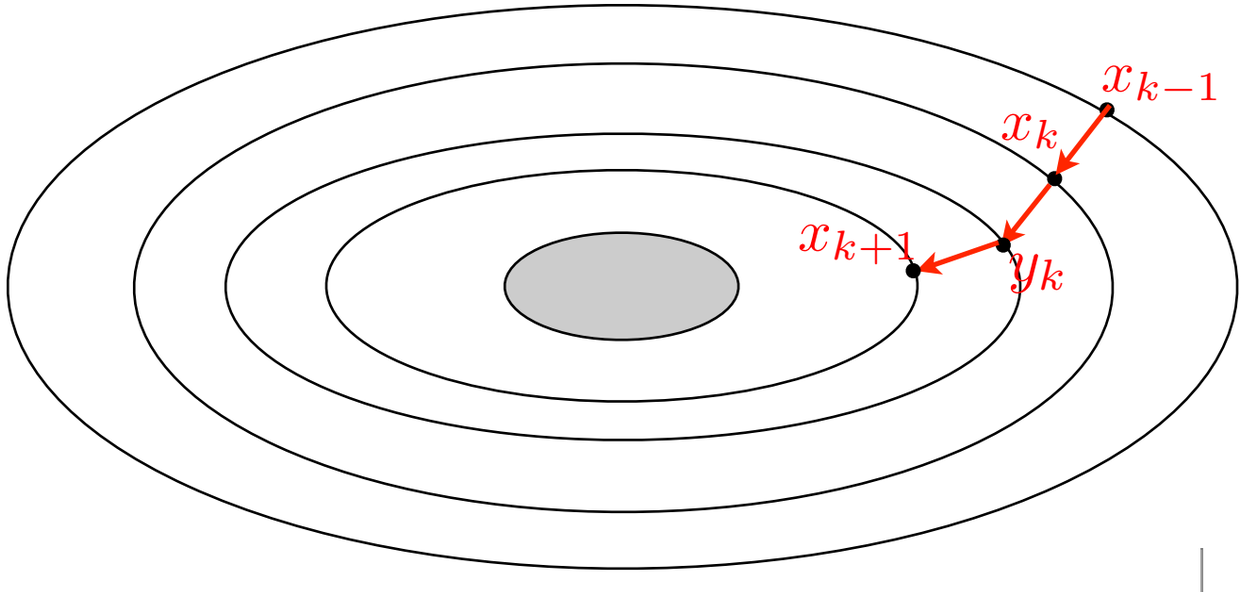




Gradient descent



Inertial gradient descent



Accelerated gradient algorithm: smooth convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \text{ and convex}$$

Theorem *Suppose that* $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$*, bounded from below and convex, that* $\mathrm{Argmin}\ (f) \neq \emptyset$*,* $\gamma \in]0, 1/L]$ *and* $\alpha \geq 3$*. Then the sequence* $(x_k)_{k\in\mathbb{N}}$ *generated by the Nesterov gradient algorithm obeys for* $k \geq \alpha - 1$

$$f(x_k) - \min f \leq \frac{(f(x_0) - \min f) + \frac{1}{2}\mathrm{dist}(x_0, \mathrm{Argmin}\ (f))^2}{(k-1)^2} \quad \textit{and} \quad \sum_{k\in\mathbb{N}} k \left\|\nabla f(x_k)\right\|^2 < +\infty.$$

Accelerated gradient algorithm: smooth convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \text{ and convex}$$

Theorem *Suppose that* $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$*, bounded from below and convex, that* $\mathrm{Argmin}\ (f) \neq \emptyset$*,* $\gamma \in]0, 1/L]$ *and* $\alpha \geq 3$*. Then the sequence* $(x_k)_{k\in\mathbb{N}}$ *generated by the Nesterov gradient algorithm obeys for* $k \geq \alpha - 1$

$$f(x_k) - \min f \leq \frac{(f(x_0) - \min f) + \frac{1}{2}\mathrm{dist}(x_0, \mathrm{Argmin}\ (f))^2}{(k-1)^2} \quad \textit{and} \quad \sum_{k\in\mathbb{N}} k \left\|\nabla f(x_k)\right\|^2 < +\infty.$$

- In view of Theorem S74, the Nesterov accelerated gradient algorithm achieves the optimal rate $O(1/k^2)$ on f.
- This means that needs $k \geq C\varepsilon^{-1/2}$ for some constant $C > 0$ to achieve precision ε on function values $f \Rightarrow$ at least $O(\varepsilon^{-1/2})$ gradient evaluations.
- For $\alpha > 3$, one can show that the rate on f is actually $o(1/k^2)$ and that the sequence $(x_k)_{k\in\mathbb{N}}$ converges to a minimizer of f. We omit the details here.

Accelerated gradient algorithm: smooth convex

Proof: Our proof is based on a Lyapunov analysis. Define $\alpha_k \overset{\text{def}}{=} 1 - \frac{\alpha}{k}$ and $t_{k+1} \overset{\text{def}}{=} \frac{k}{\alpha-1}$. It is easy to see that $t_k = 1 + t_{k+1}\alpha_k$. Given $x^\star \in \operatorname{Argmin}(f)$, we define the sequence

$$V_k \overset{\text{def}}{=} t_k^2(f(x_k) - f(x^\star)) + \frac{1}{2\gamma}\|v_k\|^2 \quad \text{and} \quad v_k \overset{\text{def}}{=} (x_{k-1} - x^\star) + t_k\,(x_k - x_{k-1})\,.$$

V_k is a non-negative sequence. We will show again that it is decreasing.

Since $\gamma \leq \frac{1}{L}$ and $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$, we have for all $x, y \in \mathbb{R}^d$

(Descent lemma S41) $$f(y - \gamma\nabla f(y)) \leq f(y) - \frac{\gamma}{2}(2 - L\gamma)\,\|\nabla f(y)\|^2 \leq f(y) - \frac{\gamma}{2}\,\|\nabla f(y)\|^2$$

(Theorem in S49) $$\leq f(x) + \langle \nabla f(y), y - x\rangle - \frac{1}{2L}\,\|\nabla f(x) - \nabla f(y)\|^2 - \frac{\gamma}{2}\,\|\nabla f(y)\|^2$$

$$\leq f(x) + \langle \nabla f(y), y - x\rangle - \frac{\gamma}{2}\,\|\nabla f(x) - \nabla f(y)\|^2 - \frac{\gamma}{2}\,\|\nabla f(y)\|^2\,. \tag{1}$$

Let us apply (1) successively at $y = y_k$ and $x = x_k$, then at $y = y_k$, $x = x^\star$. According to $x_{k+1} = y_k - \gamma\nabla f(y_k)$ and $\nabla f(x^\star) = 0$, we get

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(y_k), y_k - x_k\rangle - \frac{\gamma}{2}\,\|\nabla f(y_k)\|^2 - \frac{\gamma}{2}\,\|\nabla f(x_k) - \nabla f(y_k)\|^2 \tag{2}$$

$$f(x_{k+1}) \leq f(x^\star) + \langle \nabla f(y_k), y_k - x^\star\rangle - \frac{\gamma}{2}\|\nabla f(y_k)\|^2 - \frac{\gamma}{2}\,\|\nabla f(y_k)\|^2\,. \tag{3}$$

Multiplying (2) by $t_{k+1} - 1$, and noting that the latter is non-negative for $k \geq \alpha - 1$, then adding (3), we derive that

$$\begin{aligned} &t_{k+1}(f(x_{k+1}) - f(x^\star)) \leq (t_{k+1} - 1)(f(x_k) - f(x^\star)) + \langle \nabla f(y_k), (t_{k+1} - 1)(y_k - x_k) + y_k - x^\star\rangle \\ &- \frac{\gamma}{2}t_{k+1}\,\|\nabla f(y_k)\|^2 - \frac{\gamma}{2}(t_{k+1} - 1)\,\|\nabla f(x_k) - \nabla f(y_k)\|^2 - \frac{\gamma}{2}\,\|\nabla f(y_k)\|^2 \\ &\leq (t_{k+1} - 1)(f(x_k) - f(x^\star)) + \langle \nabla f(y_k), (t_{k+1} - 1)(y_k - x_k) + y_k - x^\star\rangle - \frac{\gamma}{2}t_{k+1}\,\|\nabla f(y_k)\|^2 - \frac{\gamma}{2}\,\|\nabla f(y_k)\|^2\,. \end{aligned} \tag{4}$$

Let us multiply (4) by t_{k+1} to make appear V_{k+1}. We obtain

$$\begin{aligned} t_{k+1}^2(f(x_{k+1}) - f(x^\star)) &\leq (t_{k+1}^2 - t_{k+1})(f(x_k) - f(x^\star)) \\ &+ t_{k+1}\langle \nabla f(y_k), (t_{k+1} - 1)(y_k - x_k) + y_k - x^\star\rangle - \frac{\gamma}{2}t_{k+1}^2\,\|\nabla f(y_k)\|^2 - \frac{\gamma}{2}t_{k+1}\,\|\nabla f(y_k)\|^2\,. \end{aligned} \tag{5}$$

Since $\alpha \geq 3$, one can check that $t_{k+1}^2 - t_{k+1} \leq t_k^2$, and (5) becomes

$$\begin{aligned} t_{k+1}^2(f(x_{k+1} - f(x^\star)) &\leq t_k^2(f(x_k) - f(x^\star)) \\ &+ t_{k+1}\langle \nabla f(y_k), (t_{k+1} - 1)(y_k - x_k) + y_k - x^\star\rangle - \frac{\gamma}{2}t_{k+1}^2\,\|\nabla f(y_k)\|^2 - \frac{\gamma}{2}t_{k+1}\,\|\nabla f(y_k)\|^2\,. \end{aligned} \tag{6}$$

Accelerated gradient algorithm: smooth convex

Proof: According to the definition of V_k, (6) reads

$$V_{k+1} - V_k \leq t_{k+1}\langle \nabla f(y_k), (t_{k+1}-1)(y_k - x_k) + y_k - x^\star\rangle - \frac{\gamma}{2} t_{k+1}^2 \left\|\nabla f(y_k)\right\|^2 \\ - \frac{\gamma}{2} t_{k+1} \left\|\nabla f(y_k)\right\|^2 + \frac{1}{2\gamma}\|v_{k+1}\|^2 - \frac{1}{2\gamma}\|v_k\|^2.$$

Let us compute this last expression with the help of the elementary identity

$$\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 = \langle v_{k+1} - v_k, v_{k+1}\rangle - \frac{1}{2}\|v_{k+1} - v_k\|^2.$$

By definition of v_k and $t_k - 1 = t_{k+1}\alpha_k$, we have

$$\begin{aligned} v_{k+1} - v_k &= x_k - x_{k-1} + t_{k+1}(x_{k+1} - x_k) - t_k(x_k - x_{k-1}) \\ &= t_{k+1}(x_{k+1} - x_k) - (t_k - 1)(x_k - x_{k-1}) \\ &= t_{k+1}\left(x_{k+1} - (x_k + \alpha_k(x_k - x_{k-1}))\right) = t_{k+1}\left(x_{k+1} - y_k\right) = -\gamma t_{k+1}\nabla f(y_k). \end{aligned}$$

Hence

$$\frac{1}{2\gamma}\|v_{k+1}\|^2 - \frac{1}{2\gamma}\|v_k\|^2 = -\frac{\gamma}{2} t_{k+1}^2 \|\nabla f(y_k)\|^2 - t_{k+1}\left\langle \nabla f(y_k), x_k - x^\star + t_{k+1}\left(x_{k+1} - x_k\right)\right\rangle.$$

Collecting the above results, we obtain

$$V_{k+1} - V_k \leq t_{k+1}\langle \nabla f(y_k), (t_{k+1}-1)(y_k - x_k) + y_k - x^\star\rangle - \gamma t_{k+1}^2 \left\|\nabla f(y_k)\right\|^2 \\ - t_{k+1}\left\langle \nabla f(y_k), x_k - x^\star + t_{k+1}\left(x_{k+1} - x_k\right)\right\rangle - \frac{\gamma}{2} t_{k+1} \left\|\nabla f(y_k)\right\|^2.$$

Equivalently $$V_{k+1} - V_k \leq t_{k+1}\langle \nabla f(y_k), A_k\rangle - \gamma t_{k+1}^2 \left\|\nabla f(y_k)\right\|^2 - \frac{\gamma}{2} t_{k+1} \left\|\nabla f(y_k)\right\|^2,$$

where

$$\begin{aligned} A_k &\stackrel{\text{def}}{=} (t_{k+1}-1)(y_k - x_k) + y_k - x_k - t_{k+1}\left(x_{k+1} - x_k\right) \\ &= t_{k+1}y_k - t_{k+1}x_k - t_{k+1}x_{k+1} + t_{k+1}x_k = t_{k+1}(y_k - x_{k+1}) = \gamma t_{k+1}\nabla f(y_k). \end{aligned}$$

Consequently

$$\begin{aligned} V_{k+1} - V_k &\leq \gamma t_{k+1}^2 \left\|\nabla f(y_k)\right\|^2 - \gamma t_{k+1}^2 \left\|\nabla f(y_k)\right\|^2 - \frac{\gamma}{2} t_{k+1} \left\|\nabla f(y_k)\right\|^2 \\ &= -\frac{\gamma}{2} t_{k+1} \left\|\nabla f(y_k)\right\|^2. \end{aligned}$$

Thus, $(V_k)_{k\in\mathbb{N}}$ is a decreasing sequence for $k \geq k_0 = \alpha - 1$, from which we get

$$f(x_k) - \min f \leq \frac{V_k}{t_k^2} \leq \frac{V_{k_0}}{t_k^2} = \frac{V_{k_0}(\alpha-1)^2}{(k-1)^2}.$$

Moreover, summing these inequalities, and since $t_k \sim k$, we get $\sum_{k\in\mathbb{N}} k \left\|\nabla f(y_k)\right\|^2 < +\infty$. Since from (4), we have $\sum_{k\in\mathbb{N}} t_k \left\|\nabla f(y_k) - \nabla f(x_k)\right\|^2 < +\infty$, the summability also holds at x_k thanks to Jensen's inequality. ■

Accelerated gradient algorithm: strongly smooth convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \text{ and convex}$$

Theorem *Suppose that $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ and μ-strongly convex. Consider the algorithm*

$$y_k = x_k + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\left(x_k - x_{k-1}\right)$$

$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k).$$

Then

$$\frac{\mu}{2}\left\|x_k - x^\star\right\|^2 \leq f(x_k) - \min f = O\left(\left(1-\sqrt{\mu/L}\right)^k\right).$$

Accelerated gradient algorithm: strongly smooth convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \text{ and convex}$$

Theorem *Suppose that $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ and μ-strongly convex. Consider the algorithm*

$$y_k = x_k + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\left(x_k - x_{k-1}\right)$$

$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k).$$

Then

$$\frac{\mu}{2}\left\|x_k - x^\star\right\|^2 \leq f(x_k) - \min f = O\left(\left(1-\sqrt{\mu/L}\right)^k\right).$$

- This is much better than the rate $O\left((1-\mu/L)^k\right)$ of gradient descent for badly-conditioned problems.

see S65

Accelerated gradient algorithm: strongly smooth convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \text{ and convex}$$

Theorem *Suppose that $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ and μ-strongly convex. Consider the algorithm*

$$y_k = x_k + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\left(x_k - x_{k-1}\right)$$

$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k).$$

Then

$$\frac{\mu}{2}\left\|x_k - x^\star\right\|^2 \leq f(x_k) - \min f = O\left(\left(1-\sqrt{\mu/L}\right)^k\right).$$

- This is much better than the rate $O\left((1-\mu/L)^k\right)$ of gradient descent for badly-conditioned problems.

see S65

Proof: See [Polyak 1987, Nesterov 2002]. ■

Summary of convergence rates

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

	Criterion	Gradient descent	Accelerated gradient descent
Non-convex	$\min_{i\in[k]} \|\nabla f(x_i)\|^2$	$O(1/k)$	
Non-convex $\cap$Ł$(1/2)$	f and $\mathrm{dist}(\cdot, \mathrm{Argmin}\,(f))$	$O(\exp(-\mu/L\ k))$	
Convex	f	$O(1/k)$ $(o(1/k))$	$O(1/k^2)$ $(o(1/k^2))$
Strongly convex	f and $\|x_k - x^\star\|^2$	$O(\exp(-\mu/L\ k))$	$O(\exp(-\sqrt{\mu/L}k))$

Outline

- Classes of functions.
- Toolbox on sequences.
- Deterministic smooth optimization.
- Stochastic approximation à la Robbins-Monro.
- Stochastic gradient descent: vanishing step-size.
- Stochastic gradient descent for finite sums.

Stochastic approximation

- Problem of finding zeros of $h : \mathbb{R}^d \to \mathbb{R}^d$:
  - h expensive to compute at all points.
  - Use random observations of values of h at certain points.
  - Main example here : finding critical points of $f : \mathbb{R}^d \to \mathbb{R}$ with $h = \nabla f$.
- Robbins and Monro algorithm [Robbins and Monro 1951, Duflo 1996] :

$$x_{k+1} = x_k - \gamma_k(h(x_k) + \varepsilon_k),$$

  ε_k is the random error $h(x_k)$.
- The Robbins-Monro algorithm cannot converge all the time (one has to control bias and variance of ε_k).
- Goals :
  - General sufficient conditions for convergence.
  - Modes of convergence : in mean, almost surely, on $h(x_k)$, on x_k.
  - Rates of convergences and choice of step-sizes.

Stochastic approximation in ML

- Population risk minimization :
  - Minimize $f(x) = \mathbb{E}_\xi \left[\ell(x, \xi)\right]$.
  - Use the gradients at i.i.d. observations.
- Empirical risk minimization :
  - Finite set of i.i.d. observations : $\xi_1, \cdots, \xi_n$.
  - Minimize $\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^n \ell(x, \xi_i)$.
  - Use the gradients at i.i.d. observations on batches $B_k \subset [n]$.
  - Special case of the above when the measure is discrete supported on $\xi_1, \cdots, \xi_n$.
  - The finite sum special structure opens the door to variance reduction.
- Online learning :
  - Compute update at iteration k after each new observation ξ_k has arrived.
  - Cumulative loss : $\frac{1}{k} \sum_{i=1}^k \ell(x_{i-1}, \xi_i)$.

Lyapunov function

$$x_{k+1} = x_k - \gamma_k (h(x_k) + \varepsilon_k)$$

- The Robbins-Monro algorithm cannot converge all the time.
- To analyze convergence, define a Lyapunov function $V : \mathbb{R}^d \to \mathbb{R}$:

(i) V is non-negative.

(ii) $V \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$.

(iii) Pseudogradient condition : $\exists \kappa \geq 0$ such that

$$\langle \nabla V(x), h(x) \rangle \geq \kappa \left\| \nabla V(x) \right\|^2, \quad \forall x \in \mathbb{R}^d.$$

(iv) Growth condition : $\exists \tau \geq 0$ such that

$$\left\| h(x) \right\|^2 \leq \tau \left(1 + \left\| \nabla V(x) \right\|^2 \right), \quad \forall x \in \mathbb{R}^d.$$

Lyapunov function

$$x_{k+1} = x_k - \gamma_k(h(x_k) + \varepsilon_k)$$

- The Robbins-Monro algorithm cannot converge all the time.
- To analyze convergence, define a Lyapunov function $V : \mathbb{R}^d \to \mathbb{R}$:

(i) V is non-negative.

(ii) $V \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$.

(iii) Pseudogradient condition : $\exists \kappa \geq 0$ such that

$$\langle \nabla V(x), h(x) \rangle \geq \kappa \left\| \nabla V(x) \right\|^2, \quad \forall x \in \mathbb{R}^d.$$

(iv) Growth condition : $\exists \tau \geq 0$ such that

$$\left\| h(x) \right\|^2 \leq \tau \left(1 + \left\| \nabla V(x) \right\|^2 \right), \quad \forall x \in \mathbb{R}^d.$$

Example *If $h = \nabla f$ for $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$, then $V = f - \inf f$ is a natural Lyapunov function. It is not the only one though.*

Martingale noise

$$x_{k+1} = x_k - \gamma_k (h(x_k) + \varepsilon_k)$$

- The i.i.d. assumption ε_k is not needed.
- Standard assumptions :

  (i) The distribution of ε_k depends only on $\mathcal{F}_k$, information up to iteration k, and $\mathscr{F} = (\mathcal{F}_k)_{k \in \mathbb{N}}$ is a *filtration* (recall notations and definitions in S53).
  - In ML : $\mathcal{F}_k = \sigma(x_0, \ldots, x_k, u_1, v_1, \ldots, u_k, v_k)$.

  (ii) Unbiasedness : $\mathbb{E}[\varepsilon_k | \mathcal{F}_k] = 0$ a.s.

  (iii) Variance : $\mathbb{E}\left[\|\varepsilon_k\|^2 | \mathcal{F}_k\right] = \sigma_k^2$.

- Observe that this entails that x_k is $\mathcal{F}_k$-measurable.

Convergence

$$x_{k+1} = x_k - \gamma_k (h(x_k) + \varepsilon_k)$$

Theorem *Suppose that*

$$0 \leq \inf_k \gamma_k \leq \sup_k \gamma_k < \frac{2\kappa}{\tau L}$$

and

$$\sum_{k\in\mathbb{N}} \gamma_k = +\infty, \quad \sum_{k\in\mathbb{N}} \gamma_k^2 < +\infty \quad \textit{and} \quad \sum_{k\in\mathbb{N}} (\gamma_k \sigma_k)^2 < +\infty.$$

Then $V(x_k)$ converges a.s. to a non-negative valued random variable, and $\liminf_{k\to\infty} \|\nabla V(x_k)\| = 0$ a.s.

Convergence

$$x_{k+1} = x_k - \gamma_k (h(x_k) + \varepsilon_k)$$

Theorem *Suppose that*

$$0 \leq \inf_k \gamma_k \leq \sup_k \gamma_k < \frac{2\kappa}{\tau L}$$

and

$$\sum_{k\in\mathbb{N}} \gamma_k = +\infty, \quad \sum_{k\in\mathbb{N}} \gamma_k^2 < +\infty \quad \textit{and} \quad \sum_{k\in\mathbb{N}} (\gamma_k \sigma_k)^2 < +\infty.$$

Then $V(x_k)$ converges a.s. to a non-negative valued random variable, and $\liminf_{k\to\infty} \|\nabla V(x_k)\| = 0$ a.s.

- For fixed noise variance $\sigma_k \equiv \sigma > 0$, our assumption on γ_k needs it to behave as $\gamma_k = C/k^{1/2+\delta}$, for $\delta \in]0, 1/2]$.
- Our assumptions allow for σ_k to grow but not too fast : critical limit for $\sigma_k = C' k^s$, $s < \delta$.

Convergence

Proof: Since $V \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$, we have by the descent lemma

$$\begin{aligned} V(x_{k+1}) &\leq V(x_k) + \langle \nabla V(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 \\ &= V(x_k) - \gamma_k \langle \nabla V(x_k), h(x_k) + \varepsilon_k \rangle + \frac{\gamma_k^2 L}{2} \|h(x_k) + \varepsilon_k\|^2 \\ &= V(x_k) - \gamma_k \langle \nabla V(x_k), h(x_k) + \varepsilon_k \rangle + \frac{\gamma_k^2 L}{2} \left(\|h(x_k)\|^2 + 2\langle h(x_k), \varepsilon_k \rangle + \|\varepsilon_k\|^2 \right). \end{aligned}$$

Taking the conditional expectation, we get

$$\begin{aligned} \mathbb{E}\left[V(x_{k+1}) \mid \mathcal{F}_k\right] &\leq V(x_k) - \gamma_k \langle \nabla V(x_k), h(x_k) + \mathbb{E}\left[\varepsilon_k \mid \mathcal{F}_k\right] \rangle \\ &\quad + \frac{\gamma_k^2 L}{2} \left(\|h(x_k)\|^2 + 2\langle h(x_k), \mathbb{E}\left[\varepsilon_k \mid \mathcal{F}_k\right] \rangle + \mathbb{E}\left[\|\varepsilon_k\|^2 \mid \mathcal{F}_k \right] \right) \\ \text{(Unbiasedness S87: zero-mean noise)} \quad &= V(x_k) - \gamma_k \langle \nabla V(x_k), h(x_k) \rangle + \frac{\gamma_k^2 L}{2} \left(\|h(x_k)\|^2 + \sigma_k^2 \right) \\ \text{(Growth condition S86)} \quad &\leq V(x_k) - \gamma_k \kappa \|\nabla V(x_k)\|^2 + \frac{\gamma_k^2 L \tau}{2} \|\nabla V(x_k)\|^2 + \frac{\gamma_k^2 L}{2} \left(\tau + \sigma_k^2 \right) \\ &= V(x_k) - \gamma_k \left(\kappa - \gamma_k \tau L/2 \right) \|\nabla V(x_k)\|^2 + \frac{\gamma_k^2 L}{2} \left(\tau + \sigma_k^2 \right). \end{aligned} \tag{1}$$

Let $\beta = \sup_k \kappa - \gamma_k \tau L/2$. We have $\beta > 0$ by assumption on γ_k. We are now in position to apply the Robbins-Siegmund lemma in S55 to get the claim.

■

Convergence rate

$$x_{k+1} = x_k - \gamma_k (h(x_k) + \varepsilon_k)$$

Theorem *Suppose that* $\sigma \overset{\text{def}}{=} \sup_k \sigma_k < +\infty$. *Choose* $\gamma_i = \frac{c}{\sqrt{k+1}}$, $i = 0, \ldots, k$, *where* $c < \frac{2\kappa}{\tau L}$. *Then*

$$\min_{i \in [k]} \left\| \mathbb{E}\left[\nabla V(x_i)\right] \right\|^2 \leq \frac{\mathbb{E}[V(x_0)] - \inf V + c^2 \frac{L\left(\tau + \sigma^2\right)}{2}}{c\beta\sqrt{k+1}}.$$

If $\gamma_k = c/\sqrt{k+1}$, *then for all* $k \in \mathbb{N}$

$$\min_{i \in [k]} \left\| \mathbb{E}\left[\nabla V(x_i)\right] \right\|^2 = O\left(\frac{\log(k+1)}{\sqrt{k+1}}\right).$$

Convergence rate

$$x_{k+1} = x_k - \gamma_k (h(x_k) + \varepsilon_k)$$

Theorem *Suppose that* $\sigma \overset{\text{def}}{=} \sup_k \sigma_k < +\infty$. *Choose* $\gamma_i = \frac{c}{\sqrt{k+1}}$, $i = 0, \ldots, k$, *where* $c < \frac{2\kappa}{\tau L}$. *Then*

$$\min_{i \in [k]} \left\| \mathbb{E}\left[\nabla V(x_i)\right] \right\|^2 \leq \frac{\mathbb{E}[V(x_0)] - \inf V + c^2 \frac{L\left(\tau + \sigma^2\right)}{2}}{c\beta\sqrt{k+1}}.$$

If $\gamma_k = c/\sqrt{k+1}$, *then for all* $k \in \mathbb{N}$

$$\min_{i \in [k]} \left\| \mathbb{E}\left[\nabla V(x_i)\right] \right\|^2 = O\left(\frac{\log(k+1)}{\sqrt{k+1}}\right).$$

- In the first claim, the number of iterations k is fixed a priori.
- The second claim is valid for an arbitrary number of iterations k.

Convergence rate

Proof: Taking the expectation in (1) in S89, we have

(Jensen's inequality)
$$\beta\left(\sum_{i=0}^{k}\gamma_i\right)\min_{i\in[k]}\left\|\mathbb{E}\left[\nabla V(x_i)\right]\right\|^2 \le \beta\left(\sum_{i=0}^{k}\gamma_i\right)\min_{i\in[k]}\mathbb{E}\left[\left\|\nabla V(x_i)\right\|^2\right]$$
$$\le \beta\sum_{i=0}^{k}\gamma_i\mathbb{E}\left[\left\|\nabla V(x_i)\right\|^2\right]$$

(Telescopic sum)
$$\le \mathbb{E}[V(x_0)] - \mathbb{E}[V(x_{k+1})] + \frac{L\left(\tau+\sigma^2\right)}{2}\sum_{i=0}^{k}\gamma_i^2$$
$$\le \mathbb{E}[V(x_0)] - \inf V + \frac{L\left(\tau+\sigma^2\right)}{2}\sum_{i=0}^{k}\gamma_i^2.$$

Thus
$$\min_{i\in[k]}\left\|\mathbb{E}\left[\nabla V(x_i)\right]\right\|^2 \le \frac{\mathbb{E}[V(x_0)] - \inf V + \frac{L\left(\tau+\sigma^2\right)}{2}\sum_{i=0}^{k}\gamma_i^2}{\beta\sum_{i=0}^{k}\gamma_i}. \tag{2}$$

The upper-bound is a convex function of $(\gamma_i)_{i\in[k]}$, and the optimal choice is $\gamma_i = c/\sqrt{k+1}$, for some constant $c > 0$, whence we get the first claim.

Let $h(t) = c/\sqrt{t+1}$. Since h is decreasing, we have for $i = 2, 3, \ldots$

(Integral test of series)
$$\int_i^{i+1} h(t)dt \le \gamma_i \le \int_{i-1}^{i} h(t)dt.$$

In turn, for $k \ge 2$ and $s \in \{1, 2\}$
$$\int_2^{k+1} h(t)^s dt \le \sum_{i=0}^{k}\gamma_i^s - (\gamma_0^s + \gamma_1^s) \le \int_1^k h(t)^s dt.$$

Thus,
$$\sum_{i=0}^{k}\gamma_i \ge \int_2^{k+1} h(t)dt = 2c(\sqrt{k+2} - \sqrt{3}) \text{ and } \sum_{i=0}^{k}\gamma_i^2 \le 3c^2/2 + \int_1^k h(t)^2 dt \le 3c^2/2 + c^2\log(k+1).$$

Inserting this into (2), we get the result. ■

Convergence under gradient domination

$$x_{k+1} = x_k - \gamma_k(h(x_k) + \varepsilon_k)$$

Theorem *In addition to the assumptions on V in S86, suppose it also obeys the curvature condition*

$$\|\nabla V(x)\|^2 \geq 2\mu V(x), \quad \forall x \in \mathbb{R}^d, \mu > 0.$$

Assume also that $\sigma \stackrel{\text{def}}{=} \sup_k \sigma_k < +\infty$. The following holds :

(i) If $\gamma_k \equiv \gamma \in]0, \kappa/(\tau L)]$, then

$$\mathbb{E}\left[V(x_k)\right] \leq \rho^k \mathbb{E}\left[V(x_0)\right] + \frac{\gamma L\left(\tau + \sigma^2\right)}{2\kappa\mu}(1-\rho^k),$$

where $\rho = 1 - \kappa\mu\gamma$. Thus

$$\limsup_{k\to+\infty} \mathbb{E}\left[V(x_k)\right] \leq \frac{\gamma L\left(\tau + \sigma^2\right)}{\kappa\mu}.$$

(ii) Suppose that $\inf_k \gamma_k \geq 0$, $\sup_k \gamma_k < \frac{2\kappa}{\tau L}$, $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$.

(a) If $\sum_{k\in\mathbb{N}} \gamma_k^2 < +\infty$ then $V(x_k) \to 0$ a.s..

(b) If $\gamma_k \to 0$ then $\mathbb{E}\left[V(x_k)\right] \to 0$.

Convergence under gradient domination

$$x_{k+1} = x_k - \gamma_k(h(x_k) + \varepsilon_k)$$

Theorem *In addition to the assumptions on V in S86, suppose it also obeys the curvature condition*

$$\|\nabla V(x)\|^2 \geq 2\mu V(x), \quad \forall x \in \mathbb{R}^d, \mu > 0.$$

Assume also that $\sigma \stackrel{\text{def}}{=} \sup_k \sigma_k < +\infty$. The following holds :

(i) If $\gamma_k \equiv \gamma \in]0, \kappa/(\tau L)]$, then

$$\mathbb{E}\left[V(x_k)\right] \leq \rho^k \mathbb{E}\left[V(x_0)\right] + \frac{\gamma L\left(\tau + \sigma^2\right)}{2\kappa\mu}(1-\rho^k),$$

where $\rho = 1 - \kappa\mu\gamma$. Thus

$$\limsup_{k\to+\infty} \mathbb{E}\left[V(x_k)\right] \leq \frac{\gamma L\left(\tau + \sigma^2\right)}{\kappa\mu}.$$

(ii) Suppose that $\inf_k \gamma_k \geq 0$, $\sup_k \gamma_k < \frac{2\kappa}{\tau L}$, $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$.

(a) If $\sum_{k\in\mathbb{N}} \gamma_k^2 < +\infty$ then $V(x_k) \to 0$ a.s..

(b) If $\gamma_k \to 0$ then $\mathbb{E}\left[V(x_k)\right] \to 0$.

- For $V = f - \min f$, the curvature condition above is nothing but the $\text{Ł}(1/2)$ condition in S51.
- Claim (i) states that for fixed step-size one has convergence in mean to a noise-dominated region.
- Convergence to 0 (a.s. or in mean) requires varying step-sizes.

Convergence under gradient domination

Proof: We embark from (1) in S89, and use the curvature inequality to see that

$$\mathbb{E}\left[V(x_{k+1}) \mid \mathcal{F}_k\right] \leq V(x_k) - 2\mu\gamma_k\left(\kappa - \gamma_k\tau L/2\right)V(x_k) + \frac{\gamma_k^2 L}{2}\left(\tau + \sigma^2\right)$$
$$= \left(1 - 2\mu\gamma_k\left(\kappa - \gamma_k\tau L/2\right)\right)V(x_k) + \frac{\gamma_k^2 L}{2}\left(\tau + \sigma^2\right).$$

(i) For fixed step-size, let the rate function $\zeta(\gamma) \stackrel{\text{def}}{=} 1 - 2\mu\gamma\left(\kappa - \gamma\tau L/2\right)$. It is easy to verify that this is a quadratic function whose minimum is attained at $\kappa/(\tau L)$, and it is decreasing on $]0, \kappa/(\tau L)]$. On this interval, it has also the upper-bound

$$\zeta(\gamma) \leq 1 - \kappa\mu\gamma = \rho.$$

Thus, taking the full expectation in the above inequality, we write

$$\mathbb{E}\left[V(x_{k+1})\right] \leq \rho\mathbb{E}\left[V(x_k)\right] + \frac{\gamma^2 L}{2}\left(\tau + \sigma^2\right). \tag{1}$$

Let $r_k = \mathbb{E}\left[V(x_k)\right]$ and $\beta = \frac{\gamma^2 L}{2}\left(\tau + \sigma^2\right)$, we have

$$r_{k+1} \leq \rho r_k + \beta.$$

Let $\nu_k = r_k - \beta/(1-\rho)$. We have

$$\nu_{k+1} = r_{k+1} - \beta/(1-\rho) \leq \rho r_k + \beta - \beta/(1-\rho) = \rho\nu_k.$$

Iterating this inequality, we obtain

$$\nu_k \leq \rho^k\nu_0 \Rightarrow r_k \leq \rho^k\nu_0 + \beta/(1-\rho) \leq \rho^k r_0 + \beta/(1-\rho)(1-\rho^k).$$

This gives the claim.

(ii) We now set $r_k = V(x_k)$, $\alpha_k = \gamma_k\mu\kappa$ and $\beta_k = \gamma_k^2\frac{L}{2}\left(\tau + \sigma^2\right)$, and thus get

$$\mathbb{E}\left[r_{k+1} \mid \mathcal{F}_k\right] \leq (1 - \alpha_k)r_k + \beta_k.$$

We now in position to invoke the lemma in S57 to get (a) and (b) since the respective assumptions are verified under our assumptions on γ_k. ■

Convergence rate under gradient domination

$$x_{k+1} = x_k - \gamma_k(h(x_k) + \varepsilon_k)$$

Theorem *Suppose that V verifies the assumptions of S86 and the curvature condition on S92. Assume also that $\sigma \overset{\text{def}}{=} \sup_k \sigma_k < +\infty$. Choose $\gamma_k = c/k$ where $c > 0$. Then*

$$\mathbb{E}\left[V(x_k)\right] = O(k^{-1}) \qquad \text{if } \kappa\mu c > 1,$$

$$\mathbb{E}\left[V(x_k)\right] = O\left(\frac{\log(k)}{k}\right) \qquad \text{if } \kappa\mu c = 1,$$

$$\mathbb{E}\left[V(x_k)\right] = O(k^{-\kappa\mu c}) \qquad \text{if } \kappa\mu c < 1.$$

- We get sublinear convergence rates.
- The convergence speed depends on the ”conditioning” of V.
- It becomes $O(k^{-1})$ if one chooses $c > 1/(\mu\kappa)$, which necessitates the knowledge of μ and κ.
- Observe from the second Chung lemma in S61 that the rate can scale as $O(k^{-1/2-\delta})$ with the choice $\gamma_k = c/k^{1/2+\delta}$, $\delta \in]0, 1/2[$. This is strictly worse that the $O(k^{-1})$ rate but no knowledge of μ or dependence on the ”conditioning” is required.

Convergence rate under gradient domination

Proof: For k large enough, we have $\gamma_k \leq \kappa/(\tau L)$. We thus obtain from the proof (1) in S93 that

$$\mathbb{E}\left[V(x_{k+1})\right] \leq (1 - \kappa\mu\gamma_k)\mathbb{E}\left[V(x_k)\right] + \frac{\gamma_k^2 L}{2}\left(\tau + \sigma^2\right).$$

It is then sufficient to invoke the first Chung lemma in S61 to conclude. ■

Summary of convergence rates

$$x_{k+1} = x_k - \gamma_k(h(x_k) + \varepsilon_k)$$

	Robbins-Monro algorithm
V general	$O(1/\sqrt{k})$
Gradient domination (known conditioning)	$O(1/k)$
Gradient domination (unknown conditioning)	$O(1/k^{(1-s)})$, $s > 0$ arbitrarily small

Outline

- Classes of functions.
- Toolbox on sequences.
- Deterministic smooth optimization.
- Stochastic approximation à la Robbins-Monro.
- Stochastic gradient descent: vanishing step-size.
- Stochastic gradient descent for finite sums.

Stochastic Gradient Descent

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Input : step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic estimate $G_k \sim P_k$ of $\nabla f(x_k)$;

  $x_{k+1} = x_k - \gamma_k G_k$;

  $k \leftarrow k + 1$.

return x_k.

Stochastic Gradient Descent

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Input : step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic estimate $G_k \sim P_k$ of $\nabla f(x_k)$;

  $x_{k+1} = x_k - \gamma_k G_k$;

  $k \leftarrow k + 1$.

return x_k.

- Population risk minimization :
  - Minimize $f(x) = \mathbb{E}_\xi\left[\ell(x,\xi)\right]$, $\xi \sim P$.
  - Sample n iid samples $(\xi_i)_{i\in[n]}$ from P.
  - Take $G_k = \frac{1}{n}\sum_{i=1}^n \nabla\ell(x_k,\xi_i)$.
- Empirical risk minimization (special case of the above) :
  - Minimize $f(x) = \frac{1}{n}\sum_{i=1}^n \ell_i(x)$.
  - Sample a batch $B_k \subset [n]$.
  - Take $G_k = \frac{1}{|B_k|}\sum_{i\in B_k} \nabla\ell_i(x_k)$.

Stochastic Gradient Descent

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Input : step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic estimate $G_k \sim P_k$ of $\nabla f(x_k)$;

  $x_{k+1} = x_k - \gamma_k G_k$;

  $k \leftarrow k + 1$.

return x_k.

Standard assumptions

- The distribution of G_k depends only on $\mathcal{F}_k$, information up to iteration k, and $\mathscr{F} = (\mathcal{F}_k)_{k\in\mathbb{N}}$ is a *filtration* (recall notations and definitions in S53).
- Unbiasedness : $\mathbb{E}\left[G_k - \nabla f(x_k)|\mathcal{F}_k\right] = 0$ a.s.
- Variance : $\mathbb{E}\left[\left\|G_k - \nabla f(x_k)\right\|^2 |\mathcal{F}_k\right] \leq \underbrace{\sigma^2}_{\text{Absolute error}} + \delta \underbrace{\left\|\nabla f(x_k)\right\|^2}_{\text{Relative error}}$, $\delta \geq 0$ a.s.

Stochastic Gradient Descent

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Input : step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

  Sample a stochastic estimate $G_k \sim P_k$ of $\nabla f(x_k)$;

  $x_{k+1} = x_k - \gamma_k G_k$;

  $k \leftarrow k + 1$.

return x_k.

Standard assumptions

- The distribution of G_k depends only on $\mathcal{F}_k$, information up to iteration k, and $\mathscr{F} = (\mathcal{F}_k)_{k\in\mathbb{N}}$ is a *filtration* (recall notations and definitions in S53).
- Unbiasedness : $\mathbb{E}\left[G_k - \nabla f(x_k)|\mathcal{F}_k\right] = 0$ a.s.
- Variance : $\mathbb{E}\left[\|G_k - \nabla f(x_k)\|^2 |\mathcal{F}_k\right] \leq \underbrace{\sigma^2}_{\text{Absolute error}} + \delta \underbrace{\|\nabla f(x_k)\|^2}_{\text{Relative error}}$, $\delta \geq 0$ a.s.

- SGD is a special case of Robbins-Monro stochastic approximation algorithm : $h(x_k) = \nabla f(x_k)$ and $\varepsilon_k = G_k - \nabla f(x_k)$. See S84

SGD: smooth non-convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$$

Theorem *Suppose that* $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ *and bounded from below. Assume that*

$$0 \le \inf_k \gamma_k \le \sup_k \gamma_k < \frac{2}{(1+\delta)L}, \quad \sum_{k\in\mathbb{N}} \gamma_k = +\infty, \quad \textit{and} \quad \sum_{k\in\mathbb{N}} \gamma_k^2 < +\infty.$$

Then

(i) $f(x_k) - \min f$ *converges a.s. to a non-negative valued random variable.*

(ii) $\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0$ *a.s.*

(iii) Choose $\gamma_i = \frac{c}{\sqrt{k+1}}$, $i = 0, \dots, k$, *where* $c < \frac{2\kappa}{(1+\delta)L}$. *Then*

$$\min_{i\in[k]} \|\mathbb{E}\left[\nabla f(x_i)\right]\|^2 \le \frac{\mathbb{E}[f(x_0)] - \min f + c^2 \frac{L\sigma^2}{2}}{c\beta\sqrt{k+1}}.$$

If $\gamma_k = c/\sqrt{k+1}$, *then for all* $k \in \mathbb{N}$

$$\min_{i\in[k]} \|\mathbb{E}\left[\nabla f(x_i)\right]\|^2 = O\left(\frac{\log(k+1)}{\sqrt{k+1}}\right).$$

(iv) If $(x_k)_{k\in\mathbb{N}}$ *is bounded a.s. then* $\operatorname{dist}(x_k, \operatorname{Crit}(f)) \to 0$ *a.s.*

SGD: smooth non-convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$$

Theorem *Suppose that* $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ *and bounded from below. Assume that*

$$0 \leq \inf_k \gamma_k \leq \sup_k \gamma_k < \frac{2}{(1+\delta)L}, \quad \sum_{k\in\mathbb{N}} \gamma_k = +\infty, \quad \textit{and} \quad \sum_{k\in\mathbb{N}} \gamma_k^2 < +\infty.$$

Then

(i) $f(x_k) - \min f$ *converges a.s. to a non-negative valued random variable.*

(ii) $\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0$ *a.s.*

(iii) Choose $\gamma_i = \frac{c}{\sqrt{k+1}}$, $i = 0, \ldots, k$, *where* $c < \frac{2\kappa}{(1+\delta)L}$. *Then*

$$\min_{i\in[k]} \|\mathbb{E}\left[\nabla f(x_i)\right]\|^2 \leq \frac{\mathbb{E}[f(x_0)] - \min f + c^2 \frac{L\sigma^2}{2}}{c\beta\sqrt{k+1}}.$$

If $\gamma_k = c/\sqrt{k+1}$, *then for all* $k \in \mathbb{N}$

$$\min_{i\in[k]} \|\mathbb{E}\left[\nabla f(x_i)\right]\|^2 = O\left(\frac{\log(k+1)}{\sqrt{k+1}}\right).$$

(iv) If $(x_k)_{k\in\mathbb{N}}$ *is bounded a.s. then* $\operatorname{dist}(x_k, \operatorname{Crit}(f)) \to 0$ *a.s.*

- In general : one needs $k \geq C\varepsilon^{-4}$ to achieve precision ε in the average gradient norm.
- But the cost per iteration can be much smaller ; i.e. much less gradient evaluations per iteration.

SGD: smooth non-convex

Proof: The key observation is that $V = f - \min f$ is a Lyapunov functionfor SGD seen as a Robbins-Monro approximation algorithm, and verifies the conditions in S86 with $\kappa = 1$ and $\tau = 1$. We then argue as in the proof in S89 to see that

$$\begin{aligned}
\mathbb{E}\left[f(x_{k+1}) \mid \mathcal{F}_k\right] &\leq f(x_k) - \gamma_k \langle \nabla f(x_k), \mathbb{E}\left[G_k \mid \mathcal{F}_k\right]\rangle \\
&\quad + \frac{\gamma_k^2 L}{2}\left(\|\nabla f(x_k)\|^2 + 2\langle \nabla f(x_k), \mathbb{E}\left[G_k - \nabla f(x_k) \mid \mathcal{F}_k\right]\rangle + \mathbb{E}\left[\|G_k - \nabla f(x_k)\|^2 \mid \mathcal{F}_k\right]\right) \\
&\leq f(x_k) - \gamma_k \|\nabla f(x_k)\|^2 + \frac{\gamma_k^2 L}{2}\left(\|\nabla f(x_k)\|^2 + \sigma^2 + \delta \|\nabla f(x_k)\|^2\right) \\
&= f(x_k) - \gamma_k \left(1 - \gamma_k(1+\delta)L/2\right)\|\nabla f(x_k)\|^2 + \frac{\sigma^2 L}{2}\gamma_k^2.
\end{aligned}$$

Let $\beta = \sup_k 1 - \gamma_k(1+\delta)L/2$.

We have $\beta > 0$ by assumption on γ_k. We are now in position to apply the Robbins-Siegmund lemma in S55 and Lemma S59 to get claims (i)-(ii).

Claim (iii) follows from Theorem S90.

For claim (iv), we start first by showing that $\lim_{k\to+\infty} \|\nabla f(x_k)\| = 0$ a.s. For this, we use Lemma S59. Since $(x_k)_{k\in\mathbb{N}}$ is almost surely bounded, there exists $r > 0$ such that $(x_k)_{k\in\mathbb{N}} \subset \mathbb{B}_r(0)$ a.s.. Convexity of $(.)^2$ and Lipschitz continuity of ∇f entail

$$\begin{aligned}
\|\nabla f(x_k)\|^2 - \|\nabla f(x_{k+1})\|^2 &\leq 2\|\nabla f(x_k)\|\left(\|\nabla f(x_k)\| - \|\nabla f(x_{k+1})\|\right) \\
&\leq 2\left(\sup_{x\in\mathbb{B}_r(0)} \|\nabla f(x)\|\right)\|\nabla f(x_k) - \nabla f(x_{k+1})\| \\
&\leq 2L\left(\sup_{x\in\mathbb{B}_r(0)} \|\nabla f(x)\|\right)\|x_{k+1} - x_k\| \\
&= \gamma_k 2L\left(\sup_{x\in\mathbb{B}_r(0)} \|\nabla f(x)\|\right)\|G_k\|.
\end{aligned}$$

SGD: smooth non-convex

Proof:

Denote $\kappa = \left(\sup_{x\in\mathbb{B}_r(0)} \|\nabla f(x)\|\right) < +\infty$. Taking the conditional expectation on both sides we obtain

$$\begin{aligned}
\|\nabla f(x_k)\|^2 - \mathbb{E}\left[\|\nabla f(x_{k+1})\|^2 \mid \mathcal{F}_k\right] &\leq \gamma_k 2L\kappa \mathbb{E}\left[\|G_k\| \mid \mathcal{F}_k\right] \\
\text{Triangle inequality} \qquad &\leq \gamma_k 2L\kappa \left(\mathbb{E}\left[\|G_k - \nabla f(x_k)\| \mid \mathcal{F}_k\right] + \|\nabla f(x_k)\|\right) \\
\text{Jensen's inequality} \qquad &\leq \gamma_k 2L\kappa \left(\mathbb{E}\left[\|G_k - \nabla f(x_k)\|^2 \mid \mathcal{F}_k\right]^{1/2} + \kappa\right) \\
\text{Assumption on the noise variance S99} \qquad &\leq \gamma_k 2L\kappa \left((\sigma^2 + \delta\kappa)^{1/2} + \kappa\right).
\end{aligned}$$

We now use Lemma S59 with $w_k = \|\nabla f(x_k)\|^2$, $\alpha_k = \gamma_k$. Indeed, we already know that $(\alpha_k w_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathscr{F})$, α_k is not summable, and the last inequality above verifies the assumption of the lemma with $\nu = 2L\kappa\left((\sigma^2 + \delta\kappa)^{1/2} + \kappa\right)$. We thus deduce that $\nabla f(x_k) \to 0$ a.s..

Now, since $(x_k)_{k\in\mathbb{N}}$ is bounded a.s., then a.s. it has convergent subsequences. Let $\left(x_{k_j}\right)_{j\in\mathbb{N}}$ be any convergent subsequence, and $\bar{x}$ its accumulation point. Then by continuity of ∇f, we have a.s.

$$\nabla f(\bar{x}) = \lim_{j\to\infty} \nabla f(x_{k_j}) = \lim_{k\to\infty} \nabla f(x_k) = 0,$$

meaning that $\bar{x}$ is an $\mathrm{Crit}(f)$-valued random variable. From continuity of $\mathrm{dist}(\cdot, \mathrm{Crit}(f))$, we obtain

$$\lim_{j\to+\infty} \mathrm{dist}(x_{k_j}, \mathrm{Crit}(f)) = 0.$$

The limit being unique (here 0) for any a.s. convergent subsequence $\left(x_{k_j}\right)_{j\in\mathbb{N}}$ means that the whole sequence $(\mathrm{dist}(x_k, \mathrm{Crit}(f)))_{k\in\mathbb{N}}$ a.s. converges to 0. ■

SGD: smooth with Ł(1/2)

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d) \cap \text{Ł}(1/2)$$

Theorem *Suppose that* $f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d) \cap \text{Ł}(1/2)$. *The following holds :*

(i) If $\gamma_k \equiv \gamma \in]0, 1/((1+\delta)L)]$, *then*

$$\mathbb{E}\left[f(x_k) - \min f\right] \leq \rho^k \mathbb{E}\left[f(x_0) - \min f\right] + \frac{\gamma L \sigma^2}{4\mu}(1-\rho^k),$$

where $\rho = 1 - 2\mu\gamma$. *Thus*

$$\limsup_{k\to+\infty} \mathbb{E}\left[f(x_k) - \min f\right] \leq \frac{\gamma L \sigma^2}{4\mu}.$$

(ii) Suppose that $\text{Argmin}\,(f) \neq \emptyset$, $\inf_k \gamma_k \geq 0$, $\sup_k \gamma_k < \frac{2}{(1+\delta)L}$, $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$.

(a) If $\sum_{k\in\mathbb{N}} \gamma_k^2 < +\infty$ *then* $f(x_k) \to \min f$ *and* $\text{dist}(x_k, \text{Argmin}\,(f)) \to 0$ *a.s..*

(b) If $\gamma_k \to 0$ *then* $\mathbb{E}\left[f(x_k) - \min f\right] \to 0$ *and* $\mathbb{E}\left[\text{dist}(x_k, \text{Argmin}\,(f))\right] \to 0$.

(iii) Choose $\gamma_k = c/k$ *where* $2\mu c > 1$. *then*

$$\mathbb{E}\left[f(x_k) - \min f\right] = O(k^{-1}) \quad \textit{and} \quad \mathbb{E}\left[\text{dist}(x_k, \text{Argmin}\,(f))^2\right] = O(k^{-1}).$$

SGD: smooth with Ł(1/2)

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d) \cap \text{Ł}(1/2)$$

Theorem *Suppose that* $f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d) \cap \text{Ł}(1/2)$. *The following holds :*

(i) If $\gamma_k \equiv \gamma \in]0, 1/((1+\delta)L)]$, *then*

$$\mathbb{E}\left[f(x_k) - \min f\right] \leq \rho^k \mathbb{E}\left[f(x_0) - \min f\right] + \frac{\gamma L \sigma^2}{4\mu}(1-\rho^k),$$

where $\rho = 1 - 2\mu\gamma$. *Thus*

$$\limsup_{k\to+\infty} \mathbb{E}\left[f(x_k) - \min f\right] \leq \frac{\gamma L \sigma^2}{4\mu}.$$

(ii) Suppose that $\text{Argmin}\,(f) \neq \emptyset$, $\inf_k \gamma_k \geq 0$, $\sup_k \gamma_k < \frac{2}{(1+\delta)L}$, $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$.

(a) If $\sum_{k\in\mathbb{N}} \gamma_k^2 < +\infty$ *then* $f(x_k) \to \min f$ *and* $\text{dist}(x_k, \text{Argmin}\,(f)) \to 0$ *a.s..*

(b) If $\gamma_k \to 0$ *then* $\mathbb{E}\left[f(x_k) - \min f\right] \to 0$ *and* $\mathbb{E}\left[\text{dist}(x_k, \text{Argmin}\,(f))\right] \to 0$.

(iii) Choose $\gamma_k = c/k$ *where* $2\mu c > 1$. *then*

$$\mathbb{E}\left[f(x_k) - \min f\right] = O(k^{-1}) \quad \textit{and} \quad \mathbb{E}\left[\text{dist}(x_k, \text{Argmin}\,(f))^2\right] = O(k^{-1}).$$

- One needs $k \geq C\varepsilon^{-1}$ to achieve precision ε on f and $\text{dist}(\cdot, \text{Argmin}\,(f))^2$ in expectation.
- This is to be contrasted with the $C\log(\epsilon^{-1})$ complexity in the deterministic case.
- The cost of gradient evaluation per iteration can however much smaller.

SGD: smooth with Ł(1/2)

Proof: Again, key observation is that $V = f - \min f$ is a Lyapunov function for SGD seen as a Robbins-Monro approximation algorithm, and verifies the conditions in S86. $f \in$ Ł$(1/2)$ also implies that the gradient domination condition in S92 is also verified. We then argue as in the proof in S93 and S101 to see that

$$\mathbb{E}\left[f(x_{k+1}) - \min f \mid \mathcal{F}_k\right] \leq f(x_k) - \min f - 4\mu\gamma_k\left(1 - \gamma_k(1+\delta)L/2\right)\left(f(x_k) - \min f\right) + \frac{\sigma^2 L}{2}\gamma_k^2$$

$$= \left(1 - 4\mu\gamma_k\left(1 - \gamma_k(1+\delta)L/2\right)\right)\left(f(x_k) - \min f\right) + \frac{\sigma^2 L}{2}\gamma_k^2.$$

(i) For fixed step-size, let the rate function $\zeta(\gamma) \stackrel{\text{def}}{=} 1 - 4\mu\gamma\left(1 - \gamma(1+\delta)L/2\right)$. It is easy to verify that this is a quadratic function whose minimum is attained at $1/((1+\delta)L)$, and it is decreasing on $]0, 1/((1+\delta)L)]$. On this interval, it has also the upper-bound $\zeta(\gamma) \leq 1 - 2\mu\gamma$.

Thus, taking the full expectation in the above inequality, we write

$$\mathbb{E}\left[f(x_{k+1}) - \min f\right] \leq \rho\mathbb{E}\left[f(x_k) - \min f\right] + \frac{\sigma^2 L\gamma^2}{2}.$$

With the exactly the same arguments as in S93, taking $\rho = 1 - 2\mu\gamma$ and $\beta = \sigma^2 L\gamma^2/2$ we get the first claim.

(ii) We now set $r_k = f(x_k) - \min f$, $\alpha_k = 2\mu\gamma_k$ and $\beta_k = \gamma_k^2\sigma^2 L/2$, and thus get

$$\mathbb{E}\left[r_{k+1} \mid \mathcal{F}_k\right] \leq (1 - \alpha_k)r_k + \beta_k. \tag{1}$$

We now in position to invoke the lemma in S57 to get (a) and (b) since the respective assumptions are verified under our assumptions on γ_k. The claims on $\operatorname{dist}(x_k, \operatorname{Argmin}(f))$ follow from those on f since $f \in$ Ł$(1/2)$ and thus has the quadratic growth in S51.

(iii) Take the full expectation in (1) and invoke Chung lemma in S61. ■

SGD: smooth strongly convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \text{ and strongly convex}$$

Theorem *Suppose that $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ and μ-strongly convex. Then it has a unique minimizer $x^\star$ and the following holds :*

(i) If $\gamma_k \equiv \gamma \in]0, 1/((1+\delta)L)]$, then

$$\mathbb{E}\left[f(x_k) - \min f\right] \leq \rho^k \mathbb{E}\left[f(x_0) - \min f\right] + \frac{\gamma L \sigma^2}{4\mu}(1-\rho^k),$$

where $\rho = 1 - 2\mu\gamma$. Thus

$$\limsup_{k\to+\infty} \mathbb{E}\left[f(x_k) - \min f\right] \leq \frac{\gamma L \sigma^2}{4\mu}.$$

(ii) Suppose that $\inf_k \gamma_k \geq 0$, $\sup_k \gamma_k < \frac{2}{(1+\delta)L}$, $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$.

(a) If $\sum_{k\in\mathbb{N}} \gamma_k^2 < +\infty$ then $f(x_k) \to \min f$ and $\|x_k - x^\star\| \to 0$ a.s..

(b) If $\gamma_k \to 0$ then $\mathbb{E}\left[f(x_k) - \min f\right] \to 0$ and $\mathbb{E}\left[\|x_k - x^\star\|\right] \to 0$.

(iii) Choose $\gamma_k = c/k$ where $2\mu c > 1$. then

$$\mathbb{E}\left[f(x_k) - \min f\right] = O(k^{-1}) \quad \textit{and} \quad \mathbb{E}\left[\|x_k - x^\star\|^2\right] = O(k^{-1}).$$

Proof: This is just a specialization of Theorem S103 since a strongly convex function is in Ł$(1/2)$ and has a unique minimizer. ■

SGD: smooth convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d) \text{ and convex}$$

Theorem *Suppose that $f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d)$, bounded from below and convex, and that $\operatorname*{Argmin}_{\mathbb{R}^d}(f) \neq \emptyset$. Assume that*

$$0 \leq \inf_k \gamma_k \leq \sup_k \gamma_k < \frac{1}{(1+\delta)L}, \quad \sum_{k\in\mathbb{N}} \gamma_k = +\infty, \quad \textit{and} \quad \sum_{k\in\mathbb{N}} \gamma_k^2 < +\infty.$$

Then

(i) $f(x_k) \to \min f$ *a.s. at the ergodic rate*

$$\mathbb{E}\left[f(\bar{x}_k) - \min f\right] \leq \frac{\mathbb{E}\left[\operatorname{dist}(x_0, \operatorname*{Argmin}_{\mathbb{R}^d}(f))^2\right] + \sigma^2 \sum_{i=0}^k \gamma_i^2}{2\sum_{i=0}^k \gamma_i},$$

where $\bar{x}_k = \sum_{i=0}^k \gamma_i x_i / \sum_{j=0}^k \gamma_j$.

(ii) If $\gamma_k = c/\sqrt{k+1}$ *for* $c < \frac{1}{(1+\delta)L}$, *then*

$$\mathbb{E}\left[f(\bar{x}_k) - \min f\right] = O\left(\frac{\log(k+1)}{\sqrt{k+1}}\right).$$

(iii) x_k *converges a.s. to a random variable valued in* $\operatorname*{Argmin}_{\mathbb{R}^d}(f)$.

SGD: smooth convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d) \text{ and convex}$$

Theorem *Suppose that $f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d)$, bounded from below and convex, and that $\operatorname*{Argmin}_{\mathbb{R}^d}(f) \neq \emptyset$. Assume that*

$$0 \leq \inf_k \gamma_k \leq \sup_k \gamma_k < \frac{1}{(1+\delta)L}, \quad \sum_{k\in\mathbb{N}} \gamma_k = +\infty, \quad \textit{and} \quad \sum_{k\in\mathbb{N}} \gamma_k^2 < +\infty.$$

Then

(i) $f(x_k) \to \min f$ *a.s. at the ergodic rate*

$$\mathbb{E}\left[f(\bar{x}_k) - \min f\right] \leq \frac{\mathbb{E}\left[\operatorname{dist}(x_0, \operatorname*{Argmin}_{\mathbb{R}^d}(f))^2\right] + \sigma^2 \sum_{i=0}^k \gamma_i^2}{2\sum_{i=0}^k \gamma_i},$$

where $\bar{x}_k = \sum_{i=0}^k \gamma_i x_i / \sum_{j=0}^k \gamma_j$.

(ii) If $\gamma_k = c/\sqrt{k+1}$ *for* $c < \frac{1}{(1+\delta)L}$, *then*

$$\mathbb{E}\left[f(\bar{x}_k) - \min f\right] = O\left(\frac{\log(k+1)}{\sqrt{k+1}}\right).$$

(iii) x_k *converges a.s. to a random variable valued in* $\operatorname*{Argmin}_{\mathbb{R}^d}(f)$.

- One needs $k \geq C\varepsilon^{-2}$ to achieve ergodic precision ε on f in expectation.
- This is to be contrasted with the $C\varepsilon^{-1}$ complexity in the deterministic case.
- The cost of gradient evaluation per iteration can however much smaller.
- Convergence of f (in ergodic sense) to a noise dominated region if the step-size is bounded away from zero.

SGD: smooth convex

Proof: (i) Denote for short $\mathcal{S} = \operatorname*{Argmin}_{\mathbb{R}^d}(f)$. Let $x^\star \in \mathcal{S}$ be the closest vector to x_k. We have

$$\begin{aligned}
\operatorname{dist}(x_{k+1},\mathcal{S})^2 &\leq \|x_{k+1}-x^\star\|^2 = \|x_k - x^\star\|^2 - 2\gamma_k\langle G_k, x_k - x^\star\rangle + \gamma_k^2\|G_k\|^2 \\
&= \operatorname{dist}(x_k,\mathcal{S})^2 - 2\gamma_k\langle G_k - \nabla f(x_k), x_k - x^\star\rangle - 2\gamma_k\langle \nabla f(x_k), x_k - x^\star\rangle \\
&\quad + \gamma_k^2\left(\|\nabla f(x_k)\|^2 + 2\langle \nabla f(x_k), G_k - \nabla f(x_k)\rangle + \|G_k - \nabla f(x_k)\|^2\right) \\
&\leq \operatorname{dist}(x_k,\mathcal{S})^2 - 2\gamma_k\langle G_k - \nabla f(x_k), x_k - x^\star\rangle - 2\gamma_k\left(f(x_k) - \min f + \frac{1}{2L}\|\nabla f(x_k)\|^2\right) \quad \text{Co-coercivity in S49} \\
&\quad + \gamma_k^2\left(\|\nabla f(x_k)\|^2 + 2\langle \nabla f(x_k), G_k - \nabla f(x_k)\rangle + \|G_k - \nabla f(x_k)\|^2\right).
\end{aligned}$$

Taking the conditional expectation and using the assumptions on G_k in S99, we get

$$\begin{aligned}
\mathbb{E}\left[\operatorname{dist}(x_{k+1},\mathcal{S})^2 \mid \mathcal{F}_k\right] &\leq \operatorname{dist}(x_k,\mathcal{S})^2 - 2\gamma_k\left(f(x_k) - \min f + \frac{1}{2L}\|\nabla f(x_k)\|^2\right) \\
&\quad + \gamma_k^2\left(\|\nabla f(x_k)\|^2 + \sigma^2 + \delta\|\nabla f(x_k)\|^2\right) \\
&= \operatorname{dist}(x_k,\mathcal{S})^2 - 2\gamma_k(f(x_k) - \min f) - \frac{\gamma_k}{L}(1 - (1+\delta)\gamma_k L)\|\nabla f(x_k)\|^2 + \gamma_k^2\sigma^2 \\
&\leq \operatorname{dist}(x_k,\mathcal{S})^2 - 2\gamma_k(f(x_k) - \min f) + \gamma_k^2\sigma^2. \qquad (1)
\end{aligned}$$

In view of the assumptions on $(\gamma_k)_{k\in\mathbb{N}}$, applying the Robbins-Siegmund lemma in S55, we have $\liminf_{k\to+\infty} f(x_k) = \min f$ a.s.. But we already know from TheoremS100 that $f(x_k) - \min f$ converges a.s. Altogether, this means that the $\liminf$ is actually a limit. For the (ergodic) rate, we take the full expectation in (1) and use convexity of f to get

$$\begin{aligned}
\text{Jensen's inequality} \quad 2\mathbb{E}\left[f(\bar{x}_k) - \min f\right] &\leq \frac{2}{\sum_{i=0}^k \gamma_i}\sum_{i=0}^k \gamma_i \mathbb{E}\left[f(x_i) - \min f\right] \\
\text{Telescopic property in (1)} \quad &\leq \frac{1}{\sum_{i=0}^k \gamma_i}\left(\mathbb{E}\left[\operatorname{dist}(x_0,\mathcal{S})^2\right] - \mathbb{E}\left[\operatorname{dist}(x_{k+1},\mathcal{S})^2\right] + \sigma^2\sum_{i=0}^k \gamma_i^2\right) \\
&\leq \frac{\mathbb{E}\left[\operatorname{dist}(x_0,\mathcal{S})^2\right] + \sigma^2\sum_{i=0}^k \gamma_i^2}{\sum_{i=0}^k \gamma_i}.
\end{aligned}$$

(ii) We argue exactly as in S91 to bound $\sum_{i=0}^k \gamma_i^2$ from above by $C\log(k+1)$ and $\sum_{i=0}^k \gamma_i$ from below by $C'\sqrt{k+1}$

SGD: smooth convex

Proof: (iii) This is the most technical part of the proof. Let $x^\star \in \mathcal{S}$. We argue as in (i) (see (1)) to see that

$$\mathbb{E}\left[\|x_{k+1}-x^\star\|^2 \mid \mathcal{F}_k\right] \leq \|x_k-x^\star\|^2 - 2\gamma_k(f(x_k)-\min f) - \frac{\gamma_k}{L}(1-(1+\delta)\gamma_k L)\|\nabla f(x_k)\|^2 + \gamma_k^2\sigma^2$$
$$\leq \|x_k-x^\star\|^2 + \gamma_k^2\sigma^2. \tag{2}$$

Applying again the Robbins-Siegmund lemma in S55, we have there exists a set of events $\Omega_{x^\star}$ (that depends on $x^\star$) such that $\mathbb{P}(\Omega_{x^\star})=1$ and for all $\omega \in \Omega_{x^\star}$,

$$(\|x_k(\omega)-x^\star\|)_{k\in\mathbb{N}} \text{ converges.}$$

We now show that there exists a set of events independent of $x^\star$, whose probability is 1 and such that the above still holds on this set. Since $\mathbb{R}^d$ is separable, there exists a countable set $\mathcal{Z}\subseteq\mathcal{S}$, such that $\mathrm{cl}(\mathcal{Z})=\mathcal{S}$. Let $\widetilde{\Omega}=\bigcap_{z\in\mathcal{Z}}\Omega_z$. Since $\mathcal{Z}$ is countable, a union bound shows

$$\mathbb{P}(\widetilde{\Omega}) = 1-\mathbb{P}\left(\bigcup_{z\in\mathcal{Z}}\Omega_z^c\right) \geq 1-\sum_{z\in\mathcal{Z}}\mathbb{P}(\Omega_z^c)=1.$$

For arbitrary $x^\star\in\mathcal{S}$, there exists a sequence $(z_j)_{j\in\mathbb{N}}\subseteq\mathcal{Z}$ such that $z_j\to x^\star$. Thus for every $j\in\mathbb{N}$ there exists $\tau_j:\Omega_{z_j}\to\mathbb{R}_+$ such that

$$\lim_{k\to+\infty}\|x_k(\omega)-z_j\| = \tau_j(\omega), \quad \forall\omega\in\Omega_{z_j}. \tag{3}$$

Now, let $\omega\in\widetilde{\Omega}$. Since $\widetilde{\Omega}\subset\Omega_{z_j}$ for any $j\in\mathbb{N}$, and using the triangle inequality and (3), we obtain that

$$\tau_j(\omega)-\|z_j-x^\star\| \leq \liminf_{k\to+\infty}\|x_k(\omega)-x^\star\| \leq \limsup_{k\to+\infty}\|x_k(\omega)-x^\star\| \leq \tau_j(\omega)+\|z_j-x^\star\|.$$

Passing to $j\to+\infty$, we deduce

$$\limsup_{j\to+\infty}\tau_j(\omega) \leq \liminf_{k\to+\infty}\|x_k(\omega)-x^\star\| \leq \limsup_{k\to+\infty}\|x_k(\omega)-x^\star\| \leq \liminf_{j\to+\infty}\tau_j(\omega),$$

whence we deduce that $\lim_{j\to+\infty}\tau_j(\omega)$ exists on the set $\widetilde{\Omega}$ of probability 1. In turn, almost surely, $\lim_{k\to+\infty}\|x_k-x^\star\|$ exists and is equal to $\lim_{j\to+\infty}\tau_j$ for any $x^\star\in\mathcal{S}$.

In particular, this shows that that $(x_k)_{k\in\mathbb{N}}$ is bounded a.s. Let $\left(x_{k_j}\right)_{j\in\mathbb{N}}$ be any converging subsequence, and $\bar{x}$ its accumulation point. Then using claim (i), we have a.s. that

$$f(\bar{x}) = \lim_{j\to\infty} f(x_{k_j}) = \lim_{k\to\infty} f(x_k) = \min f,$$

which means that $\bar{x}$ is a random variable valued in $\mathcal{S}$. But we have shown that $(\|x_k-\bar{x}\|)_{k\in\mathbb{N}}$ is a.s. convergent, and thus $\lim_{k\to+\infty}\|x_k-\bar{x}\| = \lim_{j\to+\infty}\left\|x_{k_j}-\bar{x}\right\| = 0$, i.e. $(x_k)_{k\in\mathbb{N}}$ converges a.s. to a random variable valued in $\mathcal{S}$.

■

SGD: smooth convex pointwise rate

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \text{ and convex}$$

- Can the rate $O(\log(k)/\sqrt{k})$ be improved ?
- Can one obtain a pointwise rate rather than the ergodic one ?
- Can the step-size be fixed rather than vanishing ?

SGD: smooth convex pointwise rate

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \text{ and convex}$$

- Can the rate $O(\log(k)/\sqrt{k})$ be improved ?
- Can one obtain a pointwise rate rather than the ergodic one ?
- Can the step-size be fixed rather than vanishing ?

Theorem *Suppose that $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$, bounded from below and convex, and that $\operatorname*{Argmin}_{\mathbb{R}^d}(f) \neq \emptyset$. Assume that*

$$0 \leq \inf_k \gamma_k \leq \sup_k \gamma_k < \frac{2}{(1+\delta)L}$$

and that the absolute error in the gradient (see S99) is iteration-dependent, say σ_k^2. Then

$$\mathbb{E}\left[f(x_{k+1}) - \min f\right] \leq \frac{\mathbb{E}\left[V_0\right] + \sum_{i=0}^{k}(i+1)\sigma_i^2\left(\gamma_i + L\gamma_i^2\right)/2}{k+1}.$$

SGD: smooth convex pointwise rate

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d) \text{ and convex}$$

- Can the rate $O(\log(k)/\sqrt{k})$ be improved ?
- Can one obtain a pointwise rate rather than the ergodic one ?
- Can the step-size be fixed rather than vanishing ?

Theorem *Suppose that $f \in \mathscr{C}^{1,1}_L(\mathbb{R}^d)$, bounded from below and convex, and that $\operatorname*{Argmin}_{\mathbb{R}^d}(f) \neq \emptyset$. Assume that*

$$0 \leq \inf_k \gamma_k \leq \sup_k \gamma_k < \frac{2}{(1+\delta)L}$$

and that the absolute error in the gradient (see S99) is iteration-dependent, say σ_k^2. Then

$$\mathbb{E}\left[f(x_{k+1}) - \min f\right] \leq \frac{\mathbb{E}\left[V_0\right] + \sum_{i=0}^k (i+1)\sigma_i^2\left(\gamma_i + L\gamma_i^2\right)/2}{k+1}.$$

- Vanishing fast enough noise, constant step-size $\gamma_k \equiv \gamma \in]0, 2/(1+\delta)L[$:
  - convergence at the rate $O(1/k)$ if $\sum_{k\in\mathbb{N}} k\sigma_k^2 < +\infty$.
- Non-vanishing noise $\inf_k \sigma_k > 0$: convergence to a noise dominated region if $\gamma_k = O(1/k)$.
- For non-vanishing noise, we cannot have both convergence and non-vanishing step-size in general : except for finite sums (see next chapter).
- It is not clear what can be said about global convergence of iterates $(x_k)_{k\in\mathbb{N}}$ when the step-size is fixed.

SGD: smooth convex pointwise rate

Proof: Take any $x^\star \in \operatorname{Argmin}(f)$. The proof extends the deterministic Lyapunov analysis to the stochastic case.

Define the sequence :

$$V_k \stackrel{\text{def}}{=} k(f(x_k) - \min f) + \frac{1}{2\gamma_k} \|x_k - x^\star\|^2 .$$

This is a non-negative sequence. We have

$$\begin{aligned}
V_{k+1} - V_k &= (k+1)(f(x_{k+1}) - f(x_k)) + f(x_k) - \min f + \frac{1}{2\gamma_k}\left(\|x_k - \gamma_k G_k - x^\star\|^2 - \|x_k - x^\star\|^2\right) \\
&= (k+1)(f(x_{k+1}) - f(x_k)) + f(x_k) - \min f + \frac{1}{2\gamma_k}\left(-2\gamma_k\langle G_k, x_k - x^\star\rangle + \gamma_k^2 \|G_k\|^2\right) \\
&\leq (k+1)\left(\langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2\right) + f(x_k) - \min f + \frac{1}{2\gamma_k}\left(-2\gamma_k\langle G_k, x_k - x^\star\rangle + \gamma_k^2 \|G_k\|^2\right) \\
&\leq (k+1)\left(-\gamma_k\langle \nabla f(x_k), G_k\rangle + \frac{\gamma_k^2 L}{2}\|G_k\|^2\right) + f(x_k) - \min f + \frac{1}{2\gamma_k}\left(-2\gamma_k\langle G_k, x_k - x^\star\rangle + \gamma_k^2 \|G_k\|^2\right).
\end{aligned}$$

Taking the conditional expectation on both sides and the assumptions on G_k, we get

$$\begin{aligned}
\mathbb{E}\left[V_{k+1} \mid \mathcal{F}_k\right] - V_k &\leq (k+1)\left(-\gamma_k\langle \nabla f(x_k), \mathbb{E}\left[G_k \mid \mathcal{F}_k\right]\rangle + \frac{\gamma_k^2 L}{2}\mathbb{E}\left[\|G_k\|^2 \mid \mathcal{F}_k\right]\right) + f(x_k) - \min f \\
&+ \frac{1}{2\gamma_k}\left(-2\gamma_k\langle \mathbb{E}\left[G_k \mid \mathcal{F}_k\right], x_k - x^\star\rangle + \gamma_k^2\mathbb{E}\left[\|G_k\|^2 \mid \mathcal{F}_k\right]\right) \\
&= (k+1)\left(-\gamma_k \|\nabla f(x_k)\|^2 + \frac{\gamma_k^2 L}{2}\mathbb{E}\left[\|G_k\|^2 \mid \mathcal{F}_k\right]\right) + f(x_k) - \min f \\
&+ \frac{1}{2\gamma_k}\left(-2\gamma_k\langle \nabla f(x_k), x_k - x^\star\rangle + \gamma_k^2\mathbb{E}\left[\|G_k\|^2 \mid \mathcal{F}_k\right]\right).
\end{aligned}$$

As previously shown,

$$\mathbb{E}\left[\|G_k\|^2 \mid \mathcal{F}_k\right] = (1+\delta)\|\nabla f(x_k)\|^2 + \sigma_k^2.$$

SGD: smooth convex pointwise rate

Proof: Plugging this and denoting $\rho_k \overset{\text{def}}{=} \gamma_k(1 - \gamma_k L(1+\delta)/2)$, we get

$$\mathbb{E}\left[V_{k+1} \mid \mathcal{F}_k\right] - V_k \leq (k+1)\left(-\gamma_k \left\|\nabla f(x_k)\right\|^2 + \frac{\gamma_k^2 L}{2}(1+\delta)\left\|\nabla f(x_k)\right\|^2 + \frac{\gamma_k^2 L \sigma_k^2}{2}\right) + f(x_k) - \min f$$

$$- \left\langle \nabla f(x_k), x_k - x^\star \right\rangle + \frac{\gamma_k}{2}(1+\delta)\left\|\nabla f(x_k)\right\|^2 + \frac{\gamma_k \sigma_k^2}{2}$$

$$\leq -(k+1)\rho_k \left\|\nabla f(x_k)\right\|^2 - \frac{1}{2L}\left\|\nabla f(x_k)\right\|^2 + \frac{\gamma_k(1+\delta)}{2}\left\|\nabla f(x_k)\right\|^2 + \frac{(k+1)\gamma_k^2 L\sigma_k^2}{2} + \frac{\gamma_k \sigma_k^2}{2}.$$

Co-coercivity in S49

Since $\gamma_k(1+\delta) \leq 1/L$, and taking the full expectation, we obtain

$$\mathbb{E}\left[V_{k+1}\right] \leq \mathbb{E}\left[V_k\right] - (k+1)\rho_k \mathbb{E}\left[\left\|\nabla f(x_k)\right\|^2\right] + \frac{(k+1)\gamma_k^2 L\sigma_k^2}{2} + \frac{\gamma_k \sigma_k^2}{2} \leq \mathbb{E}\left[V_k\right] + \frac{(k+1)\gamma_k^2 L\sigma_k^2}{2} + \frac{\gamma_k \sigma_k^2}{2}$$

$$\leq \ldots \leq \mathbb{E}\left[V_0\right] + \sum_{i=0}^{k} \frac{\sigma_i^2}{2}\left(\gamma_i + L(i+1)\gamma_i^2\right).$$

From the definition of V_k, we eventually get

$$\mathbb{E}\left[f(x_{k+1}) - \min f\right] \leq \frac{\mathbb{E}\left[V_0\right] + \sum_{i=0}^{k} \frac{\sigma_i^2}{2}\left(\gamma_i + L(i+1)\gamma_i^2\right)}{k+1}.$$

■

Accelerated SGD: smooth convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \text{ and convex}$$

Input : step-size sequence $(\gamma_k)_{k\in\mathbb{N}}$, $\alpha \geq 3$, x_0, stopping rule, probability distributions $(P_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$;

Initialization : $k = 0$;

while *Stopping rule not satisfied* **do**

$y_k = x_k + (1 - \frac{\alpha}{k})(x_k - x_{k-1})$;

Sample an estimate $G_k \sim P_k$ of $\nabla f(y_k)$;

$x_{k+1} = y_k - \gamma_k G_k$;

$k \leftarrow k + 1$.

return x_k.

Accelerated SGD: smooth convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \text{ and convex}$$

Theorem *Suppose that $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$, bounded from below and convex, and that $\operatorname*{Argmin}_{\mathbb{R}^d}(f) \neq \emptyset$. Let*

$$\mathbb{E}\left[G_k - \nabla f(x_k)|\mathcal{F}_k\right] = 0 \quad \textit{and} \quad \mathbb{E}\left[\left\|G_k - \nabla f(x_k)\right\|^2\right] \stackrel{\text{def}}{=} \sigma_k^2.$$

Run the accelerated SGD with non-increasing step-sizes $\gamma_k \in]0, 1/L]$. Then

$$\mathbb{E}\left[f(x_k) - \min f\right] \leq \frac{C + (\alpha-1)\left(\sqrt{2C} + 4\sum_{i=1}^k (i-1)\gamma_i\sigma_i\right)^2}{\gamma_k(k-1)^2},$$

for some constant $C > 0$.

Accelerated SGD: smooth convex

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \text{ and convex}$$

Theorem *Suppose that $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$, bounded from below and convex, and that $\operatorname*{Argmin}_{\mathbb{R}^d}(f) \neq \emptyset$. Let*

$$\mathbb{E}\left[G_k - \nabla f(x_k)|\mathcal{F}_k\right] = 0 \quad \textit{and} \quad \mathbb{E}\left[\left\|G_k - \nabla f(x_k)\right\|^2\right] \stackrel{\text{def}}{=} \sigma_k^2.$$

Run the accelerated SGD with non-increasing step-sizes $\gamma_k \in]0, 1/L]$. Then

$$\mathbb{E}\left[f(x_k) - \min f\right] \leq \frac{C + (\alpha-1)\left(\sqrt{2C} + 4\sum_{i=1}^k (i-1)\gamma_i\sigma_i\right)^2}{\gamma_k(k-1)^2},$$

for some constant $C > 0$.

- Vanishing fast enough noise, constant step-size $\gamma_k \equiv \gamma \in]0, 1/L]$:
  - convergence at the rate $O(1/k^2)$ if $\sum_{k\in\mathbb{N}} k\sigma_k < +\infty$.
- Trade-off between decreasing rate of step-sizes, noise and convergence rate : if γ_k decreases, the noise is allowed to be larger, but the convergence rate degrades.
- For non-vanishing noise, we cannot have both convergence and non-vanishing step-size in general : except for finite sums (see next chapter).

Accelerated SGD: smooth convex

Proof: Our proof is based on a Lyapunov analysis that parallels the deterministic one in S79. Let $\alpha_k \overset{\text{def}}{=} 1 - \frac{\alpha}{k}$ and $t_{k+1} \overset{\text{def}}{=} \frac{k}{\alpha-1}$ and observe that $t_k = 1 + t_{k+1}\alpha_k$. Given $x^\star \in \operatorname{Argmin}(f)$, we define the sequence

$$V_k \overset{\text{def}}{=} \gamma_k t_k^2 (f(x_k) - f(x^\star)) + \frac{1}{2}\|v_k\|^2 \quad \text{and} \quad v_k \overset{\text{def}}{=} (x_{k-1} - x^\star) + t_k (x_k - x_{k-1}).$$

Since $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$, we have for all $x, y \in \mathbb{R}^d$

$$\begin{aligned} f(y - \gamma_k G_k) &\leq f(y) - \gamma_k \langle \nabla f(y), G_k \rangle + \frac{\gamma_k^2 L}{2} \|G_k\|^2 \\ &\leq f(x) + \langle \nabla f(y), y - x \rangle - \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 - \gamma_k \langle \nabla f(y), G_k \rangle + \frac{\gamma_k^2 L}{2} \|G_k\|^2 \\ &\leq f(x) + \langle \nabla f(y), y - x \rangle - \frac{\gamma_k}{2} \|\nabla f(x) - \nabla f(y)\|^2 - \gamma_k \langle \nabla f(y), G_k \rangle + \frac{\gamma_k^2 L}{2} \|G_k\|^2, \end{aligned} \tag{1}$$

where we used that $\gamma_k \leq 1/L$ in the last line. Let us apply (1) successively at $y = y_k$ and $x = x_k$, then at $y = y_k$, $x = x^\star$. According to $x_{k+1} = y_k - \gamma_k G_k$ and $\nabla f(x^\star) = 0$, we get

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(y_k), y_k - x_k \rangle - \frac{\gamma_k}{2} \|\nabla f(x_k) - \nabla f(y_k)\|^2 - \gamma_k \langle \nabla f(y_k), G_k \rangle + \frac{\gamma_k^2 L}{2} \|G_k\|^2 \tag{2}$$

$$f(x_{k+1}) \leq \min f + \langle \nabla f(y_k), y_k - x^\star \rangle - \frac{\gamma_k}{2} \|\nabla f(y_k)\|^2 - \gamma_k \langle \nabla f(y_k), G_k \rangle + \frac{\gamma_k^2 L}{2} \|G_k\|^2. \tag{3}$$

Taking the conditional expectation on both sides of (2) and (3), and using that $\gamma_k \leq 1/L$, we get

$$\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] \leq f(x_k) + \langle \nabla f(y_k), y_k - x_k \rangle - \frac{\gamma_k}{2} \|\nabla f(x_k) - \nabla f(y_k)\|^2 - \gamma_k \|\nabla f(y_k)\|^2 + \frac{\gamma_k}{2} \mathbb{E}\left[\|G_k\|^2 \mid \mathcal{F}_k\right] \tag{4}$$

$$\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] \leq \min f + \langle \nabla f(y_k), y_k - x^\star \rangle - \frac{\gamma_k}{2} \|\nabla f(y_k)\|^2 - \gamma_k \|\nabla f(y_k)\|^2 + \frac{\gamma_k}{2} \mathbb{E}\left[\|G_k\|^2 \mid \mathcal{F}_k\right]. \tag{5}$$

Multiplying (4) by $t_{k+1} - 1$, and noting that the latter is non-negative for $k \geq \alpha - 1$, then adding (5), we derive that

$$\begin{aligned} t_{k+1}\mathbb{E}[f(x_{k+1}) - \min f \mid \mathcal{F}_k] &\leq (t_{k+1} - 1)(f(x_k) - \min f) + \langle \nabla f(y_k), (t_{k+1} - 1)(y_k - x_k) + y_k - x^\star \rangle \\ &- \gamma_k t_{k+1} \|\nabla f(y_k)\|^2 - \frac{\gamma_k}{2}(t_{k+1} - 1) \|\nabla f(x_k) - \nabla f(y_k)\|^2 - \frac{\gamma_k}{2} \|\nabla f(y_k)\|^2 + \frac{\gamma_k t_{k+1}}{2} \mathbb{E}\left[\|G_k\|^2 \mid \mathcal{F}_k\right] \end{aligned} \tag{6}$$

Accelerated SGD: smooth convex

Proof: Let us multiply (6) by t_{k+1} to make appear V_{k+1}. We obtain

$$t_{k+1}^2\mathbb{E}\left[f(x_{k+1})-\min f \mid \mathcal{F}_k\right] \leq (t_{k+1}^2-t_{k+1})(f(x_k)-\min f)$$
$$+t_{k+1}\langle\nabla f(y_k),(t_{k+1}-1)(y_k-x_k)+y_k-x^\star\rangle-\gamma_k t_{k+1}^2\left\|\nabla f(y_k)\right\|^2-\frac{\gamma_k t_{k+1}}{2}\left\|\nabla f(y_k)\right\|^2+\frac{\gamma_k t_{k+1}^2}{2}\mathbb{E}\left[\|G_k\|^2 \mid \mathcal{F}_k\right]. \tag{7}$$

Since $\alpha \geq 3$, one can check that $t_{k+1}^2 - t_{k+1} \leq t_k^2$, and (7) becomes

$$t_{k+1}^2\mathbb{E}\left[f(x_{k+1})-\min f \mid \mathcal{F}_k\right] \leq t_k^2(f(x_k)-\min f)$$
$$+t_{k+1}\langle\nabla f(y_k),(t_{k+1}-1)(y_k-x_k)+y_k-x^\star\rangle-\gamma_k t_{k+1}^2\left\|\nabla f(y_k)\right\|^2-\frac{\gamma_k t_{k+1}}{2}\left\|\nabla f(y_k)\right\|^2+\frac{\gamma_k t_{k+1}^2}{2}\mathbb{E}\left[\|G_k\|^2 \mid \mathcal{F}_k\right]. \tag{8}$$

Multiplying both sides by γ_k, γ_k that is non-increasing and according to the definition of V_k, (8) reads

$$\mathbb{E}\left[V_{k+1} \mid \mathcal{F}_k\right]-V_k \leq \gamma_k t_{k+1}\langle\nabla f(y_k),(t_{k+1}-1)(y_k-x_k)+y_k-x^\star\rangle-\gamma_k^2 t_{k+1}^2\left\|\nabla f(y_k)\right\|^2-\frac{\gamma_k^2 t_{k+1}}{2}\left\|\nabla f(y_k)\right\|^2$$
$$+\frac{1}{2}\mathbb{E}\left[\|v_{k+1}\|^2 \mid \mathcal{F}_k\right]-\frac{1}{2}\|v_k\|^2+\frac{\gamma_k^2 t_{k+1}^2}{2}\mathbb{E}\left[\|G_k\|^2 \mid \mathcal{F}_k\right]. \tag{9}$$

Arguing as in the deterministic case, we have

$$v_{k+1}-v_k = t_{k+1}\left(x_{k+1}-y_k\right) = -\gamma_k t_{k+1} G_k,$$

and hence

$$\mathbb{E}\left[\frac{1}{2}\|v_{k+1}\|^2 \mid \mathcal{F}_k\right]-\frac{1}{2}\|v_k\|^2 = \langle v_{k+1}-v_k, v_{k+1}\rangle-\frac{1}{2}\|v_{k+1}-v_k\|^2$$
$$=-\frac{\gamma_k^2 t_{k+1}^2}{2}\mathbb{E}\left[\|G_k\|^2 \mid \mathcal{F}_k\right]-\gamma_k t_{k+1}\mathbb{E}\left[\langle G_k-\nabla f(y_k), v_{k+1}\rangle \mid \mathcal{F}_k\right]$$
$$-\gamma_k t_{k+1}\left\langle\nabla f(y_k),\mathbb{E}\left[v_{k+1} \mid \mathcal{F}_k\right]\right\rangle. \tag{10}$$

Accelerated SGD: smooth convex

Proof: Inserting (10) into (9), we get

$$\mathbb{E}\left[V_{k+1} \mid \mathcal{F}_k\right] - V_k \leq \gamma_k t_{k+1} \langle \nabla f(y_k), \mathbb{E}\left[A_k \mid \mathcal{F}_k\right] \rangle - \gamma_k^2 t_{k+1}^2 \left\|\nabla f(y_k)\right\|^2 - \frac{\gamma_k^2 t_{k+1}}{2} \left\|\nabla f(y_k)\right\|^2 - \gamma_k t_{k+1} \mathbb{E}\left[\langle G_k - \nabla f(y_k), v_{k+1} \rangle \mid \mathcal{F}_k\right],$$

where

$$\begin{aligned} A_k &= (t_{k+1} - 1)(y_k - x_k) + y_k - x^\star - v_{k+1} \\ &= (t_{k+1} - 1)(y_k - x_k) + y_k - x_k - t_{k+1}\left(x_{k+1} - x_k\right) \\ &= t_{k+1} y_k - t_{k+1} x_k - t_{k+1} x_{k+1} + t_{k+1} x_k = t_{k+1}(y_k - x_{k+1}) = \gamma_k t_{k+1} G_k. \end{aligned}$$

Thus $\mathbb{E}\left[A_k \mid \mathcal{F}_k\right] = \gamma_k t_{k+1} \nabla f(y_k)$ (unbiasedness) and we arrive at

$$\begin{aligned} \mathbb{E}\left[V_{k+1} \mid \mathcal{F}_k\right] - V_k &\leq -\frac{\gamma_k^2}{2} t_{k+1} \left\|\nabla f(y_k)\right\|^2 - \gamma_k t_{k+1} \mathbb{E}\left[\langle G_k - \nabla f(y_k), v_{k+1} \rangle \mid \mathcal{F}_k\right] \\ &\leq \gamma_k t_{k+1} \mathbb{E}\left[\left\|G_k - \nabla f(y_k)\right\| \left\|v_{k+1}\right\| \mid \mathcal{F}_k\right]. \end{aligned} \tag{11}$$

Taking the full expectation in (11) and using Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \mathbb{E}\left[V_{k+1}\right] &\leq \mathbb{E}\left[V_k\right] + \gamma_k t_{k+1} \mathbb{E}\left[\left\|G_k - \nabla f(y_k)\right\|^2\right]^{1/2} \mathbb{E}\left[\left\|v_{k+1}\right\|^2\right]^{1/2} \\ &\leq \mathbb{E}\left[V_k\right] + \gamma_k t_{k+1} \sigma_k \mathbb{E}\left[\left\|v_{k+1}\right\|^2\right]^{1/2} \leq \cdots \leq \mathbb{E}\left[V_{k_0}\right] + \sum_{i=1}^{k} \gamma_i t_{i+1} \sigma_i \mathbb{E}\left[\left\|v_{i+1}\right\|^2\right]^{1/2}. \end{aligned} \tag{12}$$

where $k_0 = \lfloor \alpha - 1 \rfloor$. Observe that $t_{k+1}^2 - t_{k+1} \leq t_k^2$ implies that

$$(t_{k+1} + t_k)(t_{k+1} - t_k) = t_{k+1}^2 - t_k^2 \leq t_{k+1} \Rightarrow t_{k+1} \leq t_{k+1} - t_k \leq \frac{t_{k+1}}{t_{k+1} + t_k} \leq 1 \Rightarrow t_{k+1} \leq t_k + 1 \leq 2t_k \quad \forall k \geq 1.$$

Thus (12) becomes

$$\mathbb{E}\left[V_{k+1}\right] \leq \mathbb{E}\left[V_{k_0}\right] + 2 \sum_{i=1}^{k} \gamma_i t_i \sigma_i \mathbb{E}\left[\left\|v_{i+1}\right\|^2\right]^{1/2}. \tag{13}$$

As $V_k \geq \frac{1}{2}\left\|v_k\right\|^2$ by definition, and $\gamma_k \leq 1/L$, we obtain

$$\mathbb{E}\left[\left\|v_k\right\|^2\right] \leq 2\mathbb{E}\left[V_{k_0}\right] + 4 \sum_{i=1}^{k} \gamma_i t_i \sigma_i \mathbb{E}\left[\left\|v_i\right\|^2\right]^{1/2}$$

We can now invoke a discrete version of Gronwall's lemma (to be stated and proved shortly) to infer that

$$\mathbb{E}\left[\left\|v_k\right\|^2\right]^{1/2} \leq \sqrt{2\mathbb{E}\left[V_{k_0}\right]} + 4 \sum_{i=1}^{k} \gamma_i t_i \sigma_i, \quad \forall k \in \mathbb{N}.$$

Returning to (13), we get

$$\begin{aligned} s_k t_k^2 \mathbb{E}\left[f(x_k) - \min f\right] \leq \mathbb{E}\left[V_k\right] &\leq \mathbb{E}\left[V_{k_0}\right] + 2 \sum_{i=1}^{k} \gamma_i t_i \sigma_i \left(\sqrt{2\mathbb{E}\left[V_{k_0}\right]} + 4 \sum_{j=1}^{i} \gamma_j t_j \sigma_j\right) \\ &\leq \mathbb{E}\left[V_{k_0}\right] + \left(\sqrt{2\mathbb{E}\left[V_{k_0}\right]} + 4 \sum_{i=1}^{k} \gamma_i t_i \sigma_i\right)^2. \end{aligned}$$

Discrete Gronwall-Bellman's lemma

Lemma *Let $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$ be sequences of positive real numbers, and c is a positive real number such that*

$$a_k^2 \leq c + \sum_{i=1}^{k} b_i a_i.$$

Then

$$a_k \leq \sqrt{c} + \sum_{i=1}^{k} b_i.$$

Proof: Set $A_k \stackrel{\text{def}}{=} \sup_{1\leq j\leq k} a_j$. The, for $1 \leq l \leq k$

$$a_l^2 \leq c + A_k \sum_{i=1}^{l} b_i \leq c + A_k \sum_{i=1}^{k} b_i.$$

Passing to the supremum with respect to l, with $1 \leq l \leq k$, we obtain

$$A_k^2 \leq c + A_k \sum_{i=1}^{k} b_i.$$

This quadratic polynomial has two roots, only one of which is non-negative, and the polynomial is positive for

$$A_k \leq \frac{\sum_{i=1}^{k} b_i + \sqrt{\left(\sum_{i=1}^{k} b_i\right)^2 + 4c}}{2} \leq \sum_{i=1}^{k} b_i + \sqrt{c}.$$

■

Summary of convergence rates

$$\min_{x\in\mathbb{R}^d} f(x), \quad f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Vanishing step-size ($O(1/\sqrt{k+1})$)

	Criterion	SGD
Non-convex	$\min_{i\in[k]} \left\|\mathbb{E}\left[\nabla f(x_i)\right]\right\|^2$	$O(\log(k)/\sqrt{k})$
Non-convex $\cap$Ł$(1/2)$	$\mathbb{E}\left[f\right]$ and $\mathbb{E}\left[\mathrm{dist}(\cdot, \mathrm{Argmin}\ (f))^2\right]$	$O(1/k)$
Convex	f, ergodic	$O(\log(k)/\sqrt{k})$
Strongly convex	$\mathbb{E}\left[f\right]$ and $\mathbb{E}\left[\left\|x_k - x^\star\right\|^2\right]$	$O(1/k)$

Vanishing noise, f convex

	Criterion	Condition on the noise	Rate
SGD	$\mathbb{E}[f]$	$\left(k\sigma_k^2\right)_{k\in\mathbb{N}} \in \ell_1^+$	$O(1/k)$
Accelerated SGD	$\mathbb{E}[f]$	$\left(k\sigma_k\right)_{k\in\mathbb{N}} \in \ell_1^+$	$O(1/k^2)$

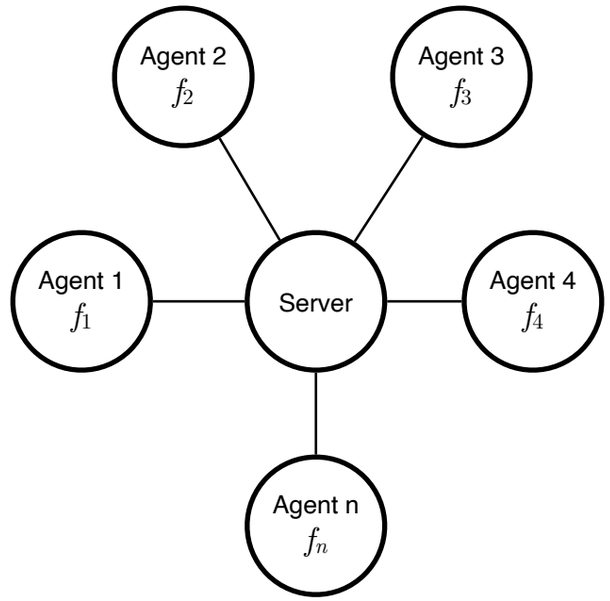
Outline

- Classes of functions.
- Toolbox on sequences.
- Deterministic smooth optimization.
- Stochastic approximation à la Robbins-Monro.
- Stochastic gradient descent: vanishing step-size.
- Stochastic gradient descent for finite sums.

Finite sums

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

- Can model empirical risk minimization $f(x) = \frac{1}{n}\sum_{i=1}^{n} \ell_i((u_i, v_i), x)$.
- Also motivated by the increasing need for distributed optimization in ML, e.g. Federated Learning :
  - Each component function f_i associated with an agent i.
  - Agents (vertices) connected through a distributed network (graph).
  - Typical graph topology : star graph where agents are connected to one central server (data privacy, agents behave independently, etc.).
  - We will not elaborate more on FL but the tools to analyze it are similar to those developed in this course.



SGD for finite sums

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Input : step-sizes $\gamma_k > 0$, minibatch size b;

Initialization : x_0;

for $k = 0, 1, \ldots$ **do**

  Uniformly randomly draw minibatch $I_k \subset [n]$ (with replacement) of size b;

  Compute gradient estimate $G_k = \frac{1}{b}\sum_{i\in I_k} \nabla f_i(x_k)$;

  $x_{k+1} = x_k - \gamma_k G_k$;

return x_k.

SGD for finite sums

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Input : step-sizes $\gamma_k > 0$, minibatch size b ;

Initialization : x_0 ;

for $k = 0, 1, \ldots$ **do**

  Uniformly randomly draw minibatch $I_k \subset [n]$ (with replacement) of size b ;

  Compute gradient estimate $G_k = \frac{1}{b}\sum_{i\in I_k} \nabla f_i(x_k)$;

  $x_{k+1} = x_k - \gamma_k G_k$;

return x_k.

- Unbiased estimate : since we sample with replacement the indices

$$\mathbb{E}\left[G_k|\mathcal{F}_k\right] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_k) = \nabla f(x_k). \quad \text{Unbiasedness assumption in S99 verified.}$$

- Variance : again sampling with replacement implies

$$\begin{aligned}\mathbb{E}\left[\left\|G_k - \nabla f(x_k)\right\|^2 |\mathcal{F}_k\right] &= \frac{1}{b^2}\mathbb{E}\left[\left\|\sum_{i\in I_k} \nabla f_i(x_k) - \nabla f(x_k)\right\|^2 |\mathcal{F}_k\right] \\ &= \frac{1}{nb}\sum_{i=1}^{n} \left\|\nabla f_i(x_k) - \nabla f(x_k)\right\|^2 \leq \frac{1}{nb}\sum_{i=1}^{n} \left\|\nabla f_i(x_k)\right\|^2 .\end{aligned}$$

If x_k is bounded a.s. or the f_i's have D-bounded gradients (e.g. logistic regression), then

$$\mathbb{E}\left[\left\|G_k - \nabla f(x_k)\right\|^2 |\mathcal{F}_k\right] \leq \frac{D}{b}.$$

Variance assumption in S99 verified.
b: Variance-complexity trade-off.

SGD for finite sums

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Input : step-sizes $\gamma_k > 0$, minibatch size b ;

Initialization : x_0 ;

for $k = 0, 1, \ldots$ **do**

  Uniformly randomly draw minibatch $I_k \subset [n]$ (with replacement) of size b ;

  Compute gradient estimate $G_k = \frac{1}{b}\sum_{i\in I_k} \nabla f_i(x_k)$;

  $x_{k+1} = x_k - \gamma_k G_k$;

return x_k.

- Unbiased estimate : since we sample with replacement the indices

$$\mathbb{E}\left[G_k|\mathcal{F}_k\right] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_k) = \nabla f(x_k). \quad \text{Unbiasedness assumption in S99 verified.}$$

- Variance : again sampling with replacement implies

$$\begin{aligned}\mathbb{E}\left[\left\|G_k - \nabla f(x_k)\right\|^2 |\mathcal{F}_k\right] &= \frac{1}{b^2}\mathbb{E}\left[\left\|\sum_{i\in I_k} \nabla f_i(x_k) - \nabla f(x_k)\right\|^2 |\mathcal{F}_k\right] \\ &= \frac{1}{nb}\sum_{i=1}^{n} \left\|\nabla f_i(x_k) - \nabla f(x_k)\right\|^2 \leq \frac{1}{nb}\sum_{i=1}^{n} \left\|\nabla f_i(x_k)\right\|^2.\end{aligned}$$

If x_k is bounded a.s. or the f_i's have D-bounded gradients (e.g. logistic regression), then

$$\mathbb{E}\left[\left\|G_k - \nabla f(x_k)\right\|^2 |\mathcal{F}_k\right] \leq \frac{D}{b}.$$

Variance assumption in S99 verified.
b: Variance-complexity trade-off.

Many results proved for SGD in the previous chapter hold for finite sums

Finite sums

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Input : step-sizes $\gamma_k > 0$, minibatch size b;

Initialization : x_0;

for $k = 0, 1, \ldots$ **do**

  Uniformly randomly draw minibatch $I_k \subset [n]$ (with replacement) of size b;

  Compute gradient estimate $G_k = \frac{1}{b}\sum_{i\in I_k} \nabla f_i(x_k)$;

  $x_{k+1} = x_k - \gamma_k G_k$;

return x_k.

Finite sums

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Input : step-sizes $\gamma_k > 0$, minibatch size b;

Initialization : x_0;

for $k = 0, 1, \ldots$ **do**

  Uniformly randomly draw minibatch $I_k \subset [n]$ (with replacement) of size b;

  Compute gradient estimate $G_k = \frac{1}{b}\sum_{i\in I_k} \nabla f_i(x_k)$;

  $x_{k+1} = x_k - \gamma_k G_k$;

return x_k.

- Issues with standard SGD studied in the previous chapter:
  - Vanishing step-size allows to annihilate the noise variance (but slow convergence).
  - Non-vanishing step-size: at best convergence to a noise dominated region.

Finite sums

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Input : step-sizes $\gamma_k > 0$, minibatch size b ;

Initialization : x_0 ;

for $k = 0, 1, \ldots$ **do**

  Uniformly randomly draw minibatch $I_k \subset [n]$ (with replacement) of size b ;

  Compute gradient estimate $G_k = \frac{1}{b}\sum_{i\in I_k} \nabla f_i(x_k)$;

  $x_{k+1} = x_k - \gamma_k G_k$;

return x_k.

- Issues with standard SGD studied in the previous chapter:
  - Vanishing step-size allows to annihilate the noise variance (but slow convergence).
  - Non-vanishing step-size: at best convergence to a noise dominated region.
- What about finite sums ?
  - A (very) special structured objective.
  - Can one afford constant step-size for such structure and achieve better convergence rate: YES !

Finite sums

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Input : step-sizes $\gamma_k > 0$, minibatch size b ;

Initialization : x_0 ;

for $k = 0, 1, \ldots$ **do**

  Uniformly randomly draw minibatch $I_k \subset [n]$ (with replacement) of size b ;

  Compute gradient estimate $G_k = \frac{1}{b}\sum_{i\in I_k} \nabla f_i(x_k)$;

  $x_{k+1} = x_k - \gamma_k G_k$;

return x_k.

- Issues with standard SGD studied in the previous chapter:
  - Vanishing step-size allows to annihilate the noise variance (but slow convergence).
  - Non-vanishing step-size: at best convergence to a noise dominated region.
- What about finite sums ?
  - A (very) special structured objective.
  - Can one afford constant step-size for such structure and achieve better convergence rate: YES !
- The key is **variance reduction**:
  - Sweep incrementally (randomly) across the functions f_i and compute gradients.
  - As in SGD, in expectation, the stochastic gradient is an unbiased estimate of the full gradient;
  - Different from SGD, the variance of the stochastic gradient converges to 0.

Variance reduction

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

- Reduce variance of random variable X using another random variable Y with known expectation :

$$Z = \alpha(X - Y) + \mathbb{E}[Y].$$

- We have

$$\mathbb{E}[Z] = \alpha\mathbb{E}[X] + (1-\alpha)\mathbb{E}[Y] \text{ and } \mathrm{Var}(Z) = \alpha^2\left(\mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X,Y)\right).$$

- If $\alpha = 1$: no bias, $\alpha < 1$: potential bias but reduced variance.
- Useful if Y positively correlated to X : reduced variance.

Variance reduction

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

- Reduce variance of random variable X using another random variable Y with known expectation :

$$Z = \alpha(X - Y) + \mathbb{E}[Y].$$

- We have

$$\mathbb{E}[Z] = \alpha\mathbb{E}[X] + (1-\alpha)\mathbb{E}[Y] \text{ and } \mathrm{Var}(Z) = \alpha^2\left(\mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X,Y)\right).$$

- If $\alpha = 1$: no bias, $\alpha < 1$: potential bias but reduced variance.
- Useful if Y positively correlated to X : reduced variance.
- Application to gradient estimation : Stochastic Variance Reduction Gradient
  - Draw uniformly randomly a minibatch $I \subset [n]$ (with replacement) such that $|I| = b$.
  - $X = \frac{1}{b}\sum_{i\in I}\nabla f_i(x)$, $Y = \frac{1}{b}\sum_{i\in I}\nabla f_i(\tilde{x})$, $\alpha = 1$, with $\tilde{x}$ stored.
  - $\mathbb{E}[Y] = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\tilde{x}) = \nabla f(\tilde{x})$, i.e. full gradient at $\tilde{x}$.
  - $\mathbb{E}[Z] = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x) = \nabla f(x)$ (never computed).
  - $\mathrm{Var}(Y) = \frac{1}{b}\mathrm{Var}(\nabla f_j(\tilde{x}))$ (sampling with replacement), j drawn uniformly at random in $[n]$, where
$$\mathrm{Var}(\nabla f_i(\tilde{x})) = \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\tilde{x})\|^2 - \|\nabla f(\tilde{x})\|^2.$$
  - Observe the influence of the minibatch size b.

Stochastic Variance Reduced Gradient (SVRG)

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Input : number of epochs S, epoch length J, step-size $\gamma > 0$, minibatch size b;

Initialization : $\tilde{x}_0 = x_0$;

for $s = 0$ **to** $S-1$ **do**

  $x_{s+1,0} = \tilde{x}_s$;

  Compute/store full gradient $\tilde{g}_s = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{x}_s)$;

  for $j = 0$ **to** $J-1$ **do**

    Uniformly randomly draw minibatch $I_j \subset [n]$ (with replacement) of size b;

    Compute gradient estimate $G_{s+1,j} = \frac{1}{b}\sum_{i\in I_j}(\nabla f_i(x_{s+1,j}) - \nabla f_i(\tilde{x}_s)) + \tilde{g}_s$;

    $x_{s+1,j+1} = x_{s+1,j} - \gamma G_{s+1,j}$;

  $\tilde{x}_{s+1} = x_{s+1,J}$

return $\tilde{x}_{s+1}$.

Stochastic Variance Reduced Gradient (SVRG)

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Input : number of epochs S, epoch length J, step-size $\gamma > 0$, minibatch size b;

Initialization : $\tilde{x}_0 = x_0$;

for $s = 0$ **to** $S - 1$ **do**

  $x_{s+1,0} = \tilde{x}_s$;

  Compute/store full gradient $\tilde{g}_s = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{x}_s)$;

  for $j = 0$ **to** $J - 1$ **do**

    Uniformly randomly draw minibatch $I_j \subset [n]$ (with replacement) of size b;

    Compute gradient estimate $G_{s+1,j} = \frac{1}{b}\sum_{i\in I_j}(\nabla f_i(x_{s+1,j}) - \nabla f_i(\tilde{x}_s)) + \tilde{g}_s$;

    $x_{s+1,j+1} = x_{s+1,j} - \gamma G_{s+1,j}$;

  $\tilde{x}_{s+1} = x_{s+1,J}$

return $\tilde{x}_{s+1}$.

- One full gradient to store per epoch.
- Gradients in inner loop are not stored (but two of them in minibatches).
- Parameters: epoch length, batch size, step-size.

Other Variance Reduced Gradient methods

- SVRG with $b = 1$:

$$G_{s+1,j} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{x}_s) + \left(\nabla f_{i_j}(x_{s+1,j}) - \nabla f_{i_j}(\tilde{x}_s)\right).$$

Other Variance Reduced Gradient methods

- SVRG with $b = 1$:

$$G_{s+1,j} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{x}_s) + \left(\nabla f_{i_j}(x_{s+1,j}) - \nabla f_{i_j}(\tilde{x}_s)\right).$$

- SAG (Stochastic Average Gradient) : $y^i_{s+1,j+1} = \begin{cases} \nabla f_{i_j}(x_{s+1,j}) & i = i_j \\ y^i_{s+1,j} & o.w. \end{cases}$

$$G_{s+1,j} = \frac{1}{n}\sum_{i=1}^{n} y^i_{s+1,j} + \frac{1}{n}\left(\nabla f_{i_j}(x_{s+1,j}) - y^{i_j}_{s+1,j}\right) = \frac{1}{n}\sum_{i=1}^{n} y^i_{s+1,j+1}.$$

Other Variance Reduced Gradient methods

- SVRG with $b = 1$:

$$G_{s+1,j} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{x}_s) + \left(\nabla f_{i_j}(x_{s+1,j}) - \nabla f_{i_j}(\tilde{x}_s)\right).$$

- SAG (Stochastic Average Gradient) : $y^i_{s+1,j+1} = \begin{cases} \nabla f_{i_j}(x_{s+1,j}) & i = i_j \\ y^i_{s+1,j} & o.w. \end{cases}$

$$G_{s+1,j} = \frac{1}{n}\sum_{i=1}^{n} y^i_{s+1,j} + \frac{1}{n}\left(\nabla f_{i_j}(x_{s+1,j}) - y^{i_j}_{s+1,j}\right) = \frac{1}{n}\sum_{i=1}^{n} y^i_{s+1,j+1}.$$

- SAGA : intermediate between SAG and SVRG :

$$G_{s+1,j} = \frac{1}{n}\sum_{i=1}^{n} y^i_{s+1,j} + \left(\nabla f_{i_j}(x_{s+1,j}) - y^{i_j}_{s+1,j}\right).$$

Other Variance Reduced Gradient methods

- SVRG with $b = 1$:

$$G_{s+1,j} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{x}_s) + \left(\nabla f_{i_j}(x_{s+1,j}) - \nabla f_{i_j}(\tilde{x}_s)\right).$$

- SAG (Stochastic Average Gradient) : $y^i_{s+1,j+1} = \begin{cases} \nabla f_{i_j}(x_{s+1,j}) & i = i_j \\ y^i_{s+1,j} & o.w. \end{cases}$

$$G_{s+1,j} = \frac{1}{n}\sum_{i=1}^{n} y^i_{s+1,j} + \frac{1}{n}\left(\nabla f_{i_j}(x_{s+1,j}) - y^{i_j}_{s+1,j}\right) = \frac{1}{n}\sum_{i=1}^{n} y^i_{s+1,j+1}.$$

- SAGA : intermediate between SAG and SVRG :

$$G_{s+1,j} = \frac{1}{n}\sum_{i=1}^{n} y^i_{s+1,j} + \left(\nabla f_{i_j}(x_{s+1,j}) - y^{i_j}_{s+1,j}\right).$$

	Unbiased estimate	Without epochs	No gradient storage	Gradient eval/step
SAG	✘	✓	✘ $O(nd)$	1
SVRG	✓	✘	✓ $O(d)$	2
SAGA	✓	✓	✘ $O(nd)$	1

Other Variance Reduced Gradient methods

- SVRG with $b = 1$:

$$G_{s+1,j} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{x}_s) + \left(\nabla f_{i_j}(x_{s+1,j}) - \nabla f_{i_j}(\tilde{x}_s)\right).$$

- SAG (Stochastic Average Gradient) : $y^i_{s+1,j+1} = \begin{cases} \nabla f_{i_j}(x_{s+1,j}) & i = i_j \\ y^i_{s+1,j} & o.w. \end{cases}$

$$G_{s+1,j} = \frac{1}{n}\sum_{i=1}^{n} y^i_{s+1,j} + \frac{1}{n}\left(\nabla f_{i_j}(x_{s+1,j}) - y^{i_j}_{s+1,j}\right) = \frac{1}{n}\sum_{i=1}^{n} y^i_{s+1,j+1}.$$

- SAGA : intermediate between SAG and SVRG :

$$G_{s+1,j} = \frac{1}{n}\sum_{i=1}^{n} y^i_{s+1,j} + \left(\nabla f_{i_j}(x_{s+1,j}) - y^{i_j}_{s+1,j}\right).$$

	Unbiased estimate	Without epochs	No gradient storage	Gradient eval/step
SAG	✘	✓	✘ $O(nd)$	1
SVRG	✓	✘	✓ $O(d)$	2
SAGA	✓	✓	✘ $O(nd)$	1

We focus in the sequel on SVRG, but many statements extend to SAGA

SVRG: smooth non-convex

$$\min_{x\in\mathbb{R}^d} f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Theorem *Suppose that $f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ for all i, and f is bounded from below. Assume that $b \leq n$ and $\gamma = \eta/L$, such that $\eta \in]0,1[$*

$$\frac{4\eta^2 J^2}{b} + \eta \leq 1.$$

Then

(i) $f(\tilde{x}_s) - \min f$ *converges a.s. to a non-negative valued random variable.*

(ii) $\sum_{s\in\mathbb{N}} \|\nabla f(\tilde{x}_s)\|^2 \to 0$ *a.s.*

(iii) $\nabla f(\tilde{x}_s) \to 0$ *a.s.*

(iv) For all $k \in \mathbb{N}$

$$\min_{(i,j)\in[s]\times[J]} \mathbb{E}\left[\|\nabla f(x_{i,j})\|\right]^2 = \frac{2(f(x_0) - \min f)}{\gamma J(s+1)}.$$

(v) If $(x_{s,j})_{s,j}$ *is bounded a.s. then* $\operatorname{dist}(\tilde{x}_s, \operatorname{Crit}(f)) \to 0$ *a.s.*

SVRG: smooth non-convex

$$\min_{x\in\mathbb{R}^d} f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

Theorem *Suppose that $f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ for all i, and f is bounded from below. Assume that $b \leq n$ and $\gamma = \eta/L$, such that $\eta \in]0,1[$*

$$\frac{4\eta^2 J^2}{b} + \eta \leq 1.$$

Then

(i) $f(\tilde{x}_s) - \min f$ *converges a.s. to a non-negative valued random variable.*

(ii) $\sum_{s\in\mathbb{N}} \|\nabla f(\tilde{x}_s)\|^2 \to 0$ *a.s.*

(iii) $\nabla f(\tilde{x}_s) \to 0$ *a.s.*

(iv) For all $k \in \mathbb{N}$

$$\min_{(i,j)\in[s]\times[J]} \mathbb{E}\left[\|\nabla f(x_{i,j})\|\right]^2 = \frac{2(f(x_0) - \min f)}{\gamma J(s+1)}.$$

(v) If $(x_{s,j})_{s,j}$ *is bounded a.s. then* $\operatorname{dist}(\tilde{x}_s, \operatorname{Crit}(f)) \to 0$ *a.s.*

- $s \gtrsim 1/(\eta J \varepsilon)$ to achieve ε accuracy.
- Bigger J (larger pass over data) is better but η smaller : trade-off in the choice of (J, b, η).
- $J = \lfloor \sqrt{b} \rfloor$ and $\eta = 1/3$. Gradient complexity :
  - n (full gradient at init.) + $sn \gtrsim n/(J\varepsilon) = 1/(b^{1/2}\varepsilon)$ (full gradient at each epoch) + $sJb \gtrsim b/\varepsilon$ (incremental gradients in inner iterations).
  - If $b = 1$, $J = 1$: $\gtrsim n + 1/\varepsilon$ gradient computations.
- Setting used in practice : $J = n$ (one pass over data per epoch), $b = 1$, $\eta = 1/(3n)$. Gradient complexity :
  - n (full gradient at init.) + $sn \gtrsim n/(\eta J\varepsilon) = n/\varepsilon$ (full gradient at each epoch) + $sJb \gtrsim n/\varepsilon$ (incremental gradients in inner iterations).
  - Overall $\gtrsim n/\varepsilon$.

SVRG: smooth non-convex

Before proving the theorem, we start with the following two lemmas.

Lemma *Let $x, g \in \mathbb{R}^d$ and define*

$$x^+ = x - \gamma g.$$

Then for any $z \in \mathbb{R}^d$, the following holds

$$\left\|x^+ - z\right\|^2 \leq \|x - z\|^2 + 2\gamma\langle g, z - x^+\rangle - \left\|x^+ - x\right\|^2.$$

Proof: Recall that x^+ can also be written as the unique minimizer

(See S65) $$x^+ = \operatorname*{argmin}_{z \in \mathbb{R}^d} 2\gamma\langle g, z - x\rangle + \|z - x\|^2.$$

This objective is strongly convex and thus, for any $z \in \mathbb{R}^d$

(3rd item of Theorem S49) $$\left\|x^+ - x\right\|^2 + 2\gamma\langle g, x^+ - x\rangle \leq \|z - x\|^2 + 2\gamma\langle g, z - x\rangle - \left\|x^+ - z\right\|^2.$$

Rearranging gives the result. ■

SVRG: smooth non-convex

Lemma *Assume that $f \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$. Let $x, g, x^+ \in \mathbb{R}^d$ as in the previous lemma and define*

$$\bar{x}^+ = x - \gamma \nabla f(x).$$

Then the following holds

$$2\gamma(f(x^+) - f(x)) \leq \gamma^2 \left\|\nabla f(x) - g\right\|^2 - (1 - \gamma L) \left\|x^+ - x\right\|^2 - \gamma^2 \left\|\nabla f(x)\right\|^2.$$

Proof: Applying the first lemma above with $z = \bar{x}^+$, we have

$$\left\|x^+ - \bar{x}^+\right\|^2 \leq \left\|x - \bar{x}^+\right\|^2 + 2\gamma\left\langle g, \bar{x}^+ - x^+\right\rangle - \left\|x^+ - x\right\|^2. \tag{1}$$

On the other hand, by the descent lemma, we have

Lemma S41
$$\begin{aligned} f(x^+) &\leq f(x) + \left\langle \nabla f(x), x^+ - x\right\rangle + \frac{L}{2}\left\|x^+ - x\right\|^2 \\ &= f(x) + \left\langle \nabla f(x), x^+ - \bar{x}^+\right\rangle + \left\langle \nabla f(x), \bar{x}^+ - x\right\rangle + \frac{L}{2}\left\|x^+ - x\right\|^2 \\ \text{Definition of } \bar{x}^+ \qquad &= f(x) + \left\langle \nabla f(x), x^+ - \bar{x}^+\right\rangle - \left\|\bar{x}^+ - x\right\|^2/\gamma + \frac{L}{2}\left\|x^+ - x\right\|^2. \end{aligned}$$

Multiplying the last inequality by 2γ and adding (1), we obtain

$$\begin{aligned} 2\gamma(f(x^+) - f(x)) &\leq 2\gamma\left\langle \nabla f(x) - g, x^+ - \bar{x}^+\right\rangle - (1 - \gamma L)\left\|x^+ - x\right\|^2 - \left\|x^+ - \bar{x}^+\right\|^2 - \left\|\bar{x}^+ - x\right\|^2 \\ &\leq \gamma^2 \left\|\nabla f(x) - g\right\|^2 + \left\|x^+ - \bar{x}^+\right\|^2 - (1 - \gamma L)\left\|x^+ - x\right\|^2 - \left\|x^+ - \bar{x}^+\right\|^2 - \gamma^2 \left\|\nabla f(x)\right\|^2 \\ &= \gamma^2 \left\|\nabla f(x) - g\right\|^2 - (1 - \gamma L)\left\|x^+ - x\right\|^2 - \gamma^2 \left\|\nabla f(x)\right\|^2. \end{aligned}$$

■

SVRG: smooth non-convex

Proof: We now turn to the proof of the theorem. Let the filtration $\mathcal{F}_{s+1,j} = \sigma((x_{p,q})_{p=0,q=0}^{s+1,j})$. Applying Lemma S128 to the SRVG iterates with $x^+ = x_{s+1,j+1}$, $x = x_{s+1,j}$ and $g = G_{s+1,j} = \frac{1}{b}\sum_{i\in I_j}(\nabla f_i(x_{s+1,j}) - \nabla f_i(\tilde{x}_s)) + \nabla f(\tilde{x}_s)$, we get

$$f(x_{s+1,j+1}) \leq f(x_{s+1,j}) + \frac{\gamma}{2}\left\|\nabla f(x_{s+1,j}) - G_{s+1,j}\right\|^2 - \frac{1-\gamma L}{2\gamma}\left\|x_{s+1,j+1} - x_{s+1,j}\right\|^2 - \frac{\gamma}{2}\left\|\nabla f(x_{s+1,j})\right\|^2 \tag{1}$$

As we sample with replacement, we have

$$\mathbb{E}\left[\frac{1}{b}\sum_{i\in I_j}\nabla f_i(x_{s+1,j}) \mid \mathcal{F}_{s+1,j}\right] = \nabla f(x_{s+1,j})$$

and (recall that $\tilde{x}_s = x_{s+1,0}$)

$$\mathbb{E}\left[\frac{1}{b}\sum_{i\in I_j}\nabla f_i(\tilde{x}_s) \mid \mathcal{F}_{s+1,j}\right] = \nabla f(\tilde{x}_s).$$

Denote $\zeta_{i,s+1,j} = \nabla f_i(x_{s+1,j}) - \nabla f_i(\tilde{x}_s)$ and $\xi_{s+1,j} = \frac{1}{b}\sum_{i\in I_j}\zeta_{i,s+1,j}$. We have by independence (sample with replacement)

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla f(x_{s+1,j}) - G_{s+1,j}\right\|^2 \mid \mathcal{F}_{s+1,j}\right] = \mathbb{E}\left[\left\|\xi_{s+1,j} - \mathbb{E}\left[\xi_{s+1,j} \mid \mathcal{F}_{s+1,j}\right]\right\|^2 \mid \mathcal{F}_{s+1,j}\right] \\
&= \mathrm{Var}\left[\xi_{s+1,j} \mid \mathcal{F}_{s+1,j}\right] = \frac{1}{b^2}\sum_{i=1}^{n}\mathrm{Var}\left[\zeta_{i,s+1,j}\mathbf{1}(i\in I_j) \mid \mathcal{F}_{s+1,j}\right] \leq \frac{1}{b^2}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\zeta_{i,s+1,j}\mathbf{1}(i\in I_j)\right\|^2 \mid \mathcal{F}_{s+1,j}\right] \\
&= \frac{1}{b^2}\sum_{i=1}^{n}\left\|\zeta_{i,s+1,j}\right\|^2 \Pr(i\in I_j) = \frac{1}{bn}\sum_{i=1}^{n}\left\|\zeta_{i,s+1,j}\right\|^2
\end{aligned} \tag{2}$$

since $\Pr(i\in I_j) = b/n$. On the other hand

$$\left\|\zeta_{i,s+1,j}\right\|^2 = \left\|\nabla f_i(x_{s+1,j}) - \nabla f_i(\tilde{x}_s)\right\|^2 \leq L^2\left\|x_{s+1,j} - \tilde{x}_s\right\|^2. \quad \text{(Lipschitz continuity of the gradient)} \tag{3}$$

Plugging (3) into (2) and the latter into (1) after taking expectation in (1), and since $\gamma L = \eta$ by definition, we get

$$\mathbb{E}\left[f(x_{s+1,j+1}) \mid \mathcal{F}_{s+1,j}\right] \leq f(x_{s+1,j}) + \frac{\eta L}{2b}\left\|x_{s+1,j} - \tilde{x}_s\right\|^2 - \frac{1-\eta}{2\gamma}\mathbb{E}\left[\left\|x_{s+1,j+1} - x_{s+1,j}\right\|^2 \mid \mathcal{F}_{s+1,j}\right] - \frac{\gamma}{2}\left\|\nabla f(x_{s+1,j})\right\|^2 \tag{4}$$

Define now $V_{s+1,j} = f(x_{s+1,j}) + c_j\left\|x_{s+1,j} - \tilde{x}_s\right\|^2$ which will serve for our Lyapunov analysis, where c_j is defined recursively as $c_j = c_{j+1}(1+1/J) + \frac{\eta L}{2b}$, with $c_J = 0$. We have

$$\begin{aligned}
\left\|x_{s+1,j+1} - \tilde{x}_s\right\|^2 &= \left\|x_{s+1,j} - \tilde{x}_s\right\|^2 + 2\langle x_{s+1,j+1} - x_{s+1,j}, x_{s+1,j} - \tilde{x}_s\rangle + \left\|x_{s+1,j+1} - x_{s+1,j}\right\|^2 \\
\text{Young's inequality} \qquad &= (1+1/J)\left\|x_{s+1,j} - \tilde{x}_s\right\|^2 + (1+J)\left\|x_{s+1,j+1} - x_{s+1,j}\right\|^2.
\end{aligned}$$

SVRG: smooth non-convex

Proof: [continued] Combining this with (2), we get

$$\begin{aligned}
\mathbb{E}\left[V_{s+1,j+1} \mid \mathcal{F}_{s+1,j}\right] &= \mathbb{E}\left[f(x_{s+1,j+1}) \mid \mathcal{F}_{s+1,j}\right] + c_{j+1}(1+1/J)\left\|x_{s+1,j} - \tilde{x}_s\right\|^2 + c_{j+1}(1+J)\mathbb{E}\left[\left\|x_{s+1,j+1} - x_{s+1,j}\right\|^2 \mid \mathcal{F}_{s+1,j}\right] \\
&\leq f(x_{s+1,j}) + \frac{\eta L}{2b}\left\|x_{s+1,j} - \tilde{x}_s\right\|^2 - \frac{1-\eta}{2\gamma}\mathbb{E}\left[\left\|x_{s+1,j+1} - x_{s+1,j}\right\|^2 \mid \mathcal{F}_{s+1,j}\right] \\
&- \frac{\gamma}{2}\left\|\nabla f(x_{s+1,j})\right\|^2 + c_{j+1}(1+1/J)\left\|x_{s+1,j} - \tilde{x}_s\right\|^2 + c_{j+1}(1+J)\mathbb{E}\left[\left\|x_{s+1,j+1} - x_{s+1,j}\right\|^2 \mid \mathcal{F}_{s+1,j}\right] \\
&= f(x_{s+1,j}) + \left(c_{j+1}(1+1/J) + \frac{\eta L}{2b}\right)\left\|x_{s+1,j} - \tilde{x}_s\right\|^2 \\
&+ \left(c_{j+1}(1+J) - \frac{1-\eta}{2\gamma}\right)\mathbb{E}\left[\left\|x_{s+1,j+1} - x_{s+1,j}\right\|^2 \mid \mathcal{F}_{s+1,j}\right] - \frac{\gamma}{2}\left\|\nabla f(x_{s+1,j})\right\|^2 \\
&= V_{s+1,j} + \left(c_{j+1}(1+J) + \frac{\eta}{2\gamma} - \frac{1}{2\gamma}\right)\mathbb{E}\left[\left\|x_{s+1,j+1} - x_{s+1,j}\right\|^2 \mid \mathcal{F}_{s+1,j}\right] - \frac{\gamma}{2}\left\|\nabla f(x_{s+1,j})\right\|^2.
\end{aligned}$$

We now show that under our assumption $4\eta^2 J^2/b + \eta \leq 1$, we have $c_{j+1}(1+J) + \eta/(2\gamma) \leq 1/2\gamma$. Indeed, by recursion on c_j, we have

$$c_j = \frac{\eta L}{2b}\frac{(1+1/J)^{J-j} - 1}{1/J} = \frac{\eta L J}{2b}\left((1+1/J)^{J-j} - 1\right).$$

With the standard inequality $\log(1+t) \leq t, t \geq 0$, we have $(1+1/J)^{J-j} = e^{(J-j)\log(1+1/J)} \leq e^{(J-j)/J} \leq e$, and thus

$$c_j \leq \frac{\eta L J}{2b}(e-1) \leq \frac{\eta L J}{b}.$$

Thus

$$c_{j+1}(1+J) + \eta/(2\gamma) \leq \frac{\eta L J}{b}(1+J) + L/2 \leq \frac{2\eta^2 J^2}{\gamma b} + \eta/(2\gamma) = \frac{1}{2\gamma}\left(\frac{4\eta^2 J^2}{b} + \eta\right) \leq \frac{1}{2\gamma}.$$

We has thus shown that

$$\mathbb{E}\left[V_{s+1,j+1} \mid \mathcal{F}_{s+1,j}\right] \leq V_{s+1,j} - \frac{\gamma}{2}\left\|\nabla f(x_{s+1,j})\right\|^2.$$

Iterating this inequality from $j = 0$ to $J-1$, we obtain

$$\mathbb{E}\left[V_{s+1,J} \mid \mathcal{F}_{s+1,J-1}\right] = \mathbb{E}\left[f(x_{s+1,J}) + c_J\left\|x_{s+1,J} - \tilde{x}_s\right\|^2 \mid \mathcal{F}_{s+1,J-1}\right] \leq f(x_{s+1,0}) + c_j\left\|x_{s+1,0} - \tilde{x}_s\right\|^2 - \frac{\gamma}{2}\sum_{j=0}^{J-1}\left\|\nabla f(x_{s+1,j})\right\|^2.$$

Since $x_{s+1,J} = \tilde{x}_{s+1}$, $x_{s+1,0} = \tilde{x}_s$ by the SVRG epoch update, the last inequality implies that

$$\mathbb{E}\left[f(\tilde{x}_{s+1}) \mid \tilde{\mathcal{F}}_s\right] \leq f(\tilde{x}_s) - \frac{\gamma}{2}\left\|\nabla f(\tilde{x}_s)\right\|^2$$

where $\tilde{\mathcal{F}}_s = \sigma\left(\tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_s\right) \subset \mathcal{F}_{s+1,J-1}$. We now apply the Robbins-Siegmund lemma in S55 to conclude.

SVRG: smooth with Ł(1/2)

$$\min_{x\in\mathbb{R}^d} f(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d), f \in \text{Ł}(1/2).$$

Theorem *Suppose that* $f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ *for all* i *and* $f \in$ Ł$(1/2)$ *and bounded from below. Assume that* $\operatorname{Argmin}(f) \neq \emptyset$ *and that* (b, γ, η) *are chosen according to Theorem S126. Then*

(i) $f(\tilde{x}_s) - \min f \to 0$ *and* $\operatorname{dist}(\tilde{x}_s, \operatorname{Argmin}(f)) \to 0$ *a.s.*

(ii) Moreover

$$\frac{\mu}{2}\mathbb{E}\left[\operatorname{dist}(\tilde{x}_s, \operatorname{Argmin}(f))^2\right] \leq \mathbb{E}\left[f(\tilde{x}_s) - \min f\right] = (1-\gamma\mu)^s \left(f(x_0) - \min f\right).$$

SVRG: smooth with Ł(1/2)

$$\min_{x\in\mathbb{R}^d} f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d), f \in \text{Ł}(1/2).$$

Theorem *Suppose that* $f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ *for all* i *and* $f \in$ Ł$(1/2)$ *and bounded from below. Assume that* $\operatorname{Argmin}(f) \neq \emptyset$ *and that* (b, γ, η) *are chosen according to Theorem S126. Then*

(i) $f(\tilde{x}_s) - \min f \to 0$ *and* $\operatorname{dist}(\tilde{x}_s, \operatorname{Argmin}(f)) \to 0$ *a.s.*

(ii) Moreover

$$\frac{\mu}{2}\mathbb{E}\left[\operatorname{dist}(\tilde{x}_s, \operatorname{Argmin}(f))^2\right] \leq \mathbb{E}\left[f(\tilde{x}_s) - \min f\right] = (1-\gamma\mu)^s \left(f(x_0) - \min f\right).$$

- $s \gtrsim (\mu\gamma)^{-1}\log(1/\varepsilon)$ to achieve ε accuracy.
- $J = \lfloor\sqrt{b}\rfloor$ and $\eta = 1/3$. Convergence rate $1 - \mu/(3L)$. Gradient complexity :
  - n (full gradient at init.) + $sn \gtrsim n(L/\mu)\log(1/\varepsilon)$ (full gradient at each epoch) + $sJb \gtrsim b^{3/2}(L/\mu)\log(1/\varepsilon)$ (incremental gradients in inner iterations).
  - If $b = 1$, $J = 1$: $\gtrsim n(L/\mu)\log(1/\varepsilon)$ gradient computations.
- $J = n$ (one pass over data per epoch), $b = 1$, $\eta = 1/(3n)$. Convergence rate $1 - \mu/(3nL)$. Gradient complexity :
  - n (full gradient at init.) + $sn \gtrsim n(L/\mu)\log(1/\varepsilon)$ (full gradient at each epoch) + $sJb \gtrsim n(L/\mu)\log(1/\varepsilon)$ (incremental gradients in inner iterations).
  - Overall $\gtrsim n(L/\mu)\log(1/\varepsilon)$ gradient computations.
- The first setting with $b = J = 1$ has better convergence rate while having the same gradient complexity.

SVRG: smooth with Ł$(1/2)$

Proof: From the proof in S130 we have shown that the SVRG iterates satisfy

$$\mathbb{E}\left[f(\tilde{x}_{s+1}) \mid \tilde{\mathcal{F}}_s\right] \leq f(\tilde{x}_s) - \frac{\gamma}{2}\left\|\nabla f(\tilde{x}_s)\right\|^2.$$

Using the Ł$(1/2)$ inequality in S51, we get

$$\mathbb{E}\left[f(\tilde{x}_{s+1}) \mid \tilde{\mathcal{F}}_s - \min f\right] \leq f(\tilde{x}_s)\min f - \gamma\mu(f(\tilde{x}_s)\min f) = (1-\gamma\mu)(f(\tilde{x}_s)\min f).$$

Using Lemma S57, we get claim (i). Taking the whole expectation and iterating, we get the exponential convergence in (ii). ■

SVRG: smooth convex

$$\min_{x\in\mathbb{R}^d} f(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \cap \text{convex}.$$

Theorem *Suppose that* $f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ *and convex for each* i*, and* f *is bounded from below. Assume also that* $\mathcal{S} \stackrel{\text{def}}{=} \operatorname{Argmin}(f) \neq \emptyset$ *and that* (b, γ, η) *are chosen such that*

$$\gamma = \eta/L \quad \textit{with} \quad 2\eta\frac{1+(1+\delta)J}{b} \leq 1,$$

for some $\delta > 0$*. Then,*

(i) $\sum_{s\in\mathbb{N}}\sum_{j=0}^{J-1}(f(x_{s,j}) - \min f) < +\infty$ *a.s.*

(ii) $\forall j \in \{0,\dots,J-1\}, f(x_{s,j}) - \min f \to 0$ *a.s. as* $s \to +\infty$ *at the ergodic rate*

$$\mathbb{E}\left[f(\overline{x}_s) - \min f\right] \leq \frac{bL}{4\delta\eta^2 J^2}\frac{\operatorname{dist}(x_0,\mathcal{S})^2}{s+1},$$

where $\overline{x}_s = \sum_{j=0}^{J-1}\sum_{i=0}^{s} x_{i,j}/(J(s+1))$.

(iii) $\tilde{x}_s$ *converges a.s. to an* $\mathcal{S}$*-valued random variable.*

SVRG: smooth convex

$$\min_{x\in\mathbb{R}^d} f(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d) \cap \text{convex}.$$

Theorem *Suppose that* $f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ *and convex for each* i*, and* f *is bounded from below. Assume also that* $\mathcal{S} \stackrel{\text{def}}{=} \text{Argmin}(f) \neq \emptyset$ *and that* (b, γ, η) *are chosen such that*

$$\gamma = \eta/L \quad \textit{with} \quad 2\eta\frac{1+(1+\delta)J}{b} \leq 1,$$

for some $\delta > 0$*. Then,*

(i) $\sum_{s\in\mathbb{N}}\sum_{j=0}^{J-1}(f(x_{s,j}) - \min f) < +\infty$ *a.s.*

(ii) $\forall j \in \{0,\ldots,J-1\}, f(x_{s,j}) - \min f \to 0$ *a.s. as* $s \to +\infty$ *at the ergodic rate*

$$\mathbb{E}\left[f(\overline{x}_s) - \min f\right] \leq \frac{bL}{4\delta\eta^2 J^2}\frac{\text{dist}(x_0,\mathcal{S})^2}{s+1},$$

where $\overline{x}_s = \sum_{j=0}^{J-1}\sum_{i=0}^{s} x_{i,j}/(J(s+1))$.

(iii) $\tilde{x}_s$ *converges a.s. to an* $\mathcal{S}$*-valued random variable.*

- $s \gtrsim b/(\delta\eta^2 J^2\varepsilon)$ to achieve ε accuracy.
- Larger J and smaller b but trade-off with the choice of η.
- $J = \lfloor\sqrt{b}\rfloor$, $\delta = 1/2$, $\eta = 1/5 \Rightarrow$ iteration complexity $s \gtrsim 1/\varepsilon$. Gradient complexity :
  - n (full gradient at init.) + $sn \gtrsim n/\varepsilon$ (full gradient at each epoch) + $sJb \gtrsim b^{3/2}/\varepsilon$ (incremental gradients in inner iterations).
  - Overall $\gtrsim n/\varepsilon$ gradient computations if $b \leq n^{2/3}$.
- $J = n$ (one pass over data per epoch), $b = 1$, $\delta = 1/2$, $\eta = 1/(2+3n)) \Rightarrow$ iteration complexity $s \gtrsim 1/\varepsilon$. Gradient complexity :
  - n (full gradient at init.) + $sn \gtrsim n/\varepsilon$ (full gradient at each epoch) + $sJb \gtrsim n/\varepsilon$ (incremental gradients in inner iterations).
  - Overall $\gtrsim n/\varepsilon$.

SVRG: smooth convex

Proof: Recall the result convergence theorem for the SGD in S106 and the corresponding proof. The lesson taught by that result is that if the variance vanishes at an appropriate rate, then we are done. This is precisely what we will show for SVRG thanks to variance reduction.

We adopt the same notation as in the proof of Theorem S126 and follow closely the proof in S107. Denote for short $\mathcal{S} = \operatorname*{Argmin}_{\mathbb{R}^d} (f)$. Let $x^\star \in \mathcal{S}$ be the closest vector to $x_{s+1,j}$. Recall $G_{s+1,j} = \frac{1}{b}\sum_{i\in I_j}(\nabla f_i(x_{s+1,j}) - \nabla f_i(\tilde{x}_s)) + \nabla f(\tilde{x}_s)$. Since we sample with replacement, we have (see S130) that $\mathbb{E}\left[G_{s+1,j} \mid \mathcal{F}_{s+1,j}\right] = \nabla f(x_{s+1,j})$. Thus, from (1) in S107 applied to the SVRG, we have

$$\mathbb{E}\left[\operatorname{dist}(x_{s+1,j+1},\mathcal{S})^2 \mid \mathcal{F}_{s+1,j}\right] \leq \operatorname{dist}(x_{s+1,j},\mathcal{S})^2 - 2\gamma(f(x_{s+1,j}) - \min f) - \frac{\gamma}{L}(1-\gamma L)\left\|\nabla f(x_{s+1,j})\right\|^2 + \gamma^2\mathbb{E}\left[\left\|\nabla f(x_{s+1,j}) - G_{s+1,j}\right\|^2 \mid \mathcal{F}_{s+1,j}\right]. \quad (1)$$

Let $\zeta_{i,s+1,j} = \nabla f_i(x_{s+1,j}) - \nabla f_i(\tilde{x}_s)$ and $\xi_{s+1,j} = \frac{1}{b}\sum_{i\in I_j}\zeta_{i,s+1,j}$. It is straightforward to see that $\mathbb{E}\left[\zeta_{i,s+1,j} \mid \mathcal{F}_{s+1,j}\right] = \mathbb{E}\left[\xi_{s+1,j} \mid \mathcal{F}_{s+1,j}\right] = \nabla f(x_{s+1,j}) - \nabla f_i(\tilde{x}_s)$. Thus

$$\begin{aligned}
\mathbb{E}\left[\left\|\nabla f(x_{s+1,j}) - G_{s+1,j}\right\|^2 \mid \mathcal{F}_{s+1,j}\right] &= \operatorname{Var}\left[\xi_{s+1,j} \mid \mathcal{F}_{s+1,j}\right] = \frac{1}{b^2}\sum_{i=1}^n \operatorname{Var}\left[\zeta_{i,s+1,j}\mathbf{1}(i\in I_j) \mid \mathcal{F}_{s+1,j}\right] \\
&\leq \frac{1}{b^2}\sum_{i=1}^n \mathbb{E}\left[\left\|\zeta_{i,s+1,j}\mathbf{1}(i\in I_j)\right\|^2 \mid \mathcal{F}_{s+1,j}\right] = \frac{1}{b^2}\sum_{i=1}^n \left\|\zeta_{i,s+1,j}\right\|^2 \Pr(i\in I_j) = \frac{1}{bn}\sum_{i=1}^n \left\|\zeta_{i,s+1,j}\right\|^2 \\
&= \frac{1}{bn}\sum_{i=1}^n \left\|\nabla f_i(x_{s+1,j}) - \nabla f_i(\tilde{x}_s)\right\|^2. \quad (2)
\end{aligned}$$

For the rhs term, use Jensens's inequality and Theorem S49 on $f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d)$ to get

$$\begin{aligned}
\left\|\nabla f_i(x_{s+1,j}) - \nabla f_i(\tilde{x}_s)\right\|^2 &\leq 2\left(\left\|\nabla f_i(x_{s+1,j}) - \nabla f_i(x^\star)\right\|^2 + \left\|\nabla f_i(\tilde{x}_s) - \nabla f_i(x^\star)\right\|^2\right) \\
&\leq 4L\left((f_i(x_{s+1,j}) - f_i(x^\star) - \langle\nabla f_i(x^\star), x_{s+1,j} - x^\star\rangle) + (f_i(\tilde{x}_s) - f_i(x^\star) - \langle\nabla f_i(x^\star), \tilde{x}_s - x^\star\rangle)\right).
\end{aligned}$$

Averaging this inequality over $i \in [n]$, and using that $\nabla f(x^\star) = 0$, we get

$$\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(x_{s+1,j}) - \nabla f_i(\tilde{x}_s)\right\|^2 \leq 4L\left((f(x_{s+1,j}) - \min f) + (f(\tilde{x}_s) - \min f)\right).$$

Plugging this into (2) and then in (1), rearranging and dropping the gradient term in (1) since $\eta = \gamma L < 1$, we obtain

$$\mathbb{E}\left[\operatorname{dist}(x_{s+1,j+1},\mathcal{S})^2 \mid \mathcal{F}_{s+1,j}\right] \leq \operatorname{dist}(x_{s+1,j},\mathcal{S})^2 - 2\gamma\left(1 - \frac{2\eta}{b}\right)(f(x_{s+1,j}) - \min f) + \frac{4\gamma^2 L}{b}(f(\tilde{x}_s) - \min f). \quad (3)$$

SVRG: smooth convex

Proof: Iterating (3) from $j = 0$ to $J - 1$, we obtain

$$\mathbb{E}\left[\operatorname{dist}(x_{s+1,J}, \mathcal{S})^2 \mid \mathcal{F}_{s+1,J}\right] \leq \operatorname{dist}(x_{s+1,0}, \mathcal{S})^2 - 2\gamma\left(1 - \frac{2\eta}{b}\right)\sum_{j=0}^{J-1}\left(f(x_{s+1,j}) - \min f\right) + \frac{4\gamma\eta J}{b}\left(f(\tilde{x}_s) - \min f\right)$$

$$= \operatorname{dist}(x_{s+1,0}, \mathcal{S})^2 - 2\gamma\left(1 - \frac{2\eta}{b}\right)\sum_{j=1}^{J-1}\left(f(x_{s+1,j}) - \min f\right) - \left(2\gamma\left(1 - \frac{2\eta}{b}\right) - \frac{4\gamma\eta J}{b}\right)\left(f(\tilde{x}_s) - \min f\right).$$

Under our assumption on the parameters, we have

$$2\gamma\left(1 - \frac{2\eta}{b}\right) - \frac{4\gamma\eta J}{b} \geq \frac{4\delta\gamma\eta J}{b} \text{ and } 2\gamma\left(1 - \frac{2\eta}{b}\right) \geq \frac{4(1+\delta)\gamma\eta J}{b} > \frac{4\delta\gamma\eta J}{b}.$$

Therefore,

$$\mathbb{E}\left[\operatorname{dist}(x_{s+1,J}, \mathcal{S})^2 \mid \mathcal{F}_{s+1,J}\right] \leq \operatorname{dist}(x_{s+1,0}, \mathcal{S})^2 - \frac{4\delta\gamma\eta J}{b}\sum_{j=0}^{J-1}\left(f(x_{s+1,j}) - \min f\right).$$

Since $x_{s+1,J} = \tilde{x}_{s+1}$, $x_{s+1,0} = \tilde{x}_s$ by the SVRG epoch update, the last inequality reads

$$\mathbb{E}\left[\operatorname{dist}(\tilde{x}_{s+1}, \mathcal{S})^2 \mid \mathcal{F}_{s+1,J}\right] \leq \operatorname{dist}(\tilde{x}_s, \mathcal{S})^2 - \frac{4\delta\gamma\eta J}{b}\sum_{j=0}^{J-1}\left(f(x_{s+1,j}) - \min f\right). \tag{4}$$

We are now in position to invoke the Robbins-Siegmund lemma S55 to get that $\operatorname{dist}(\tilde{x}_s, \mathcal{S})$ converges a.s. to a non-negative random variable, and that

$$\sum_{j=0}^{J-1}\left(f(x_{s+1,j}) - \min f\right) < +\infty \text{ a.s.}$$

This proves claim (i). The first part of (ii) follows from (i) since all the terms in the series are non-negative. For the rate, we take the full expectation in (4) and use Jensen's inequality to see that

$$\frac{4\delta\eta^2 J}{bL}\left(f(\overline{x}_s) - \min f\right) \leq \frac{4\delta\eta^2 J}{bL}\frac{1}{J(s+1)}\sum_{i=0}^{s}\sum_{j=0}^{J-1}\left(f(x_{i+1,j}) - \min f\right)$$

$$\leq \frac{\operatorname{dist}(x_0, \mathcal{S})^2 - \mathbb{E}[\operatorname{dist}(\tilde{x}_{s+1}, \mathcal{S})^2]}{J(s+1)} \leq \frac{\operatorname{dist}(x_0, \mathcal{S})^2}{J(s+1)}.$$

Let us prove the last claim (iii). Observe that our proof, and in particular inequality (4), remains valid replacing $\operatorname{dist}(\tilde{x}_s, \mathcal{S})^2$ by $\|\tilde{x}_s - x^\star\|^2$ for any $x^\star \in \mathcal{S}$. $f(\tilde{x}_s) \to \min f$ a.s. from (ii), and we have shown that $\|\tilde{x}_s - x^\star\|$ converges a.s. to a non-negative random variable. However, as we observed in the proof of SGD (see S108), the event of probability 1 over which convergence of $\|\tilde{x}_s - x^\star\|$ occurs depends on $x^\star$. We then follow exactly the SGD proof in S108 to conclude that there exists a set of event $\widetilde{\Omega}$ of probability one on which $\tilde{x}_s$ indeed converges to an $\mathcal{S}$-valued random variable. ■

Summary of convergence rates

$$\min_{x\in\mathbb{R}^d} f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i \in \mathscr{C}_L^{1,1}(\mathbb{R}^d).$$

	Criterion	SVRG	Cond. param.
Non-convex	$\min_{(i,j)\in[s]\times[J]} \left\|\mathbb{E}\left[\nabla f(x_{i,j})\right]\right\|^2$	$O(1/(\eta J s))$	$\gamma = \eta/L,\ \eta^2 J^2/b + \eta \leq 1$
Non-convex $\cap$Ł$(1/2)$	$\mathbb{E}\left[f\right]$ and $\mathbb{E}\left[\mathrm{dist}(\cdot, \mathrm{Argmin}\ (f))^2\right]$	$O(\exp(-\gamma/\mu s))$	Same
Convex	$\mathbb{E}\left[f\right]$, ergodic	$O(b/(\eta^2 J^2 s)$	$\gamma = \eta/L, \delta > 0, 2\eta\frac{1+(1+\delta)J}{b} \leq 1$.

Bibliography

H. Robbins and S. Monro, A stochastic approximation method. The Annals of Mathematical Statistics, Vol. 22, pp. 400–407, 1951.

S. Linnainmaa, Taylor expansion of the accumulated rounding error". BIT Numerical Mathematics, Vol. 16 No. 2, pp. 146–160, 1976.

A. Griewank and A. Walther, Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, 2nd ed., SIAM, Philadelphia, 2008.

B. T. Polyak, Introduction to optimization, Optimization Software, 1987.

M. Duflo. Algorithmes stochastiques. Springer-Verlag, 1996.

H. Robbins and D. Siegmund, A Convergence Theorem for Non Negative Almost Supermartingales and Some Applications, in Herbert Robbins Selected Papers, Springer New York, pp. 111–135, 1985.

A.S. Nemirovsky, D.B. Yudin, Problem complexity and method efficiency in optimization, John Wiley and Sons, 1983.

Y. Nesterov, Introductory lectures on convex optimization: A basic course, volume 87 of Applied Optimization. Kluwer, 2004.

Merci
Questions ?