Optimization
Part VI: Stochastic optimization

Barbara Pascal, **Nelly Pustelnik**

CNRS, Laboratoire de Physique de l'ENS de Lyon, Univ. Lyon 1
nelly.pustelnik@ens-lyon.fr

barbara.pascal@ens-lyon.fr

Motivations

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x)$$

Motivations

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x)$$

Stochastic optimization: sequence of *random* variables $(x_n)_{n \in \mathbb{N}}$ s.t.

$$f(x_n) \xrightarrow{\mathbb{E}} \inf_{x \in \mathcal{H}} f(x) \quad \text{and possibly} \quad x_n \xrightarrow{\mathbb{E}} \widehat{x}$$

Motivations

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x)$$

Stochastic optimization: sequence of *random* variables $(x_n)_{n \in \mathbb{N}}$ s.t.

$$f(x_n) \xrightarrow{\mathbb{E}} \inf_{x \in \mathcal{H}} f(x) \quad \text{and possibly} \quad x_n \xrightarrow{\mathbb{E}} \widehat{x}$$

f contains *randomness*, i.e.,

$$f(x) \equiv \mathbb{E} F(x)$$

<u>E.g.:</u> noisy observations

$$\{F_s(x), s = 1, \ldots, S\} \text{ i.i.d.}$$

Motivations

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x)$$

Stochastic optimization: sequence of *random* variables $(x_n)_{n \in \mathbb{N}}$ s.t.

$$f(x_n) \xrightarrow{\mathbb{E}} \inf_{x \in \mathcal{H}} f(x) \quad \text{and possibly} \quad x_n \xrightarrow{\mathbb{E}} \widehat{x}$$

f contains *randomness*, i.e.,

$$f(x) \equiv \mathbb{E} F(x)$$

E.g.: noisy observations

$$\{F_s(x), s = 1, \ldots, S\} \text{ i.i.d.}$$

Inject randomness

- to be tractable
- to converge faster

E.g.: sum of functions $S \gg 1$

$$f(x) = \frac{1}{S} \sum_{s=1}^{S} F_s(x)$$

Machine learning: empirical and expected risks

Example of **binary classification**:

Machine learning: empirical and expected risks

Example of **binary classification**:

- Training set $\mathcal{T} = \big\{(\boldsymbol{u}_s, z_s), \quad \boldsymbol{u}_s \in \mathbb{R}^M, z_s \in \{-1, 1\}, s = 1, \ldots, S\big\}$

Machine learning: empirical and expected risks

Example of **binary classification**:

- Training set $\mathcal{T} = \big\{(\boldsymbol{u}_s, z_s), \quad \boldsymbol{u}_s \in \mathbb{R}^M, z_s \in \{-1, 1\}, s = 1, \dots, S\big\}$
- Linear predictor model $z \sim \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u}) \rangle$, $\phi : \mathbb{R}^M \to \mathbb{R}^N$ feature map

Machine learning: empirical and expected risks

Example of **binary classification**:

- Training set $\mathcal{T} = \big\{(\boldsymbol{u}_s, z_s), \quad \boldsymbol{u}_s \in \mathbb{R}^M, z_s \in \{-1, 1\}, s = 1, \ldots, S\big\}$
- Linear predictor model $z \sim \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u})\rangle$, $\phi : \mathbb{R}^M \to \mathbb{R}^N$ feature map
- Loss $F\left(z, \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u})\rangle\right)$ measures the prediction error

Machine learning: empirical and expected risks

Example of **binary classification**:

- Training set $\mathcal{T} = \left\{(\boldsymbol{u}_s, z_s), \quad \boldsymbol{u}_s \in \mathbb{R}^M, z_s \in \{-1, 1\}, s = 1, \dots, S\right\}$
- Linear predictor model $z \sim \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u}) \rangle$, $\phi : \mathbb{R}^M \to \mathbb{R}^N$ feature map
- Loss $F\left(z, \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u}) \rangle\right)$ measures the prediction error

Empirical risk: *error on the* **training** *set*

$$\frac{1}{S} \sum_{s=1}^{S} F\left(z_s, \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u}_s) \rangle\right) = \frac{1}{S} \sum_{s=1}^{S} F_s(\boldsymbol{\vartheta})$$

Machine learning: empirical and expected risks

Example of **binary classification**:

- Training set $\mathcal{T} = \big\{(\boldsymbol{u}_s, z_s), \quad \boldsymbol{u}_s \in \mathbb{R}^M, z_s \in \{-1, 1\}, s = 1, \ldots, S\big\}$
- Linear predictor model $z \sim \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u})\rangle$, $\phi : \mathbb{R}^M \to \mathbb{R}^N$ feature map
- Loss $F\left(z, \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u})\rangle\right)$ measures the prediction error

Empirical risk: *error on the* **training** *set*

$$\frac{1}{S} \sum_{s=1}^{S} F\left(z_s, \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u}_s)\rangle\right) = \frac{1}{S} \sum_{s=1}^{S} F_s(\boldsymbol{\vartheta})$$

Expected risk or *generalization* error *error on the* **testing** *set*

$$\mathbb{E}_{(\boldsymbol{u}, z)} F\left(z, \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u})\rangle\right) = f(\boldsymbol{\vartheta})$$

Machine learning: empirical and expected risks

Example of **binary classification**:

- Training set $\mathcal{T} = \left\{(\boldsymbol{u}_s, z_s), \quad \boldsymbol{u}_s \in \mathbb{R}^M, z_s \in \{-1,1\}, s = 1,\dots,S\right\}$
- Linear predictor model $z \sim \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u})\rangle$, $\phi : \mathbb{R}^M \to \mathbb{R}^N$ feature map
- Loss $F\left(z, \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u})\rangle\right)$ measures the prediction error

Empirical risk: *error on the* **training** *set*

$$\frac{1}{S}\sum_{s=1}^{S} F\left(z_s, \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u}_s)\rangle\right) = \frac{1}{S}\sum_{s=1}^{S} F_s(\boldsymbol{\vartheta})$$

Expected risk or *generalization* error *error on the* **testing** *set*

$$\mathbb{E}_{(\boldsymbol{u},z)} F\left(z, \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u})\rangle\right) = f(\boldsymbol{\vartheta})$$

Stochastic (sub)gradient descent [Robbins-Monro]

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x) \equiv \mathbb{E} F(x)$$

Stochastic (sub)gradient descent [Robbins-Monro]

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x) \equiv \mathbb{E} F(x)$$

- no access to $\nabla f(x)$ (or to $u \in \partial f(x)$),

Stochastic (sub)gradient descent [Robbins-Monro]

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x) \equiv \mathbb{E} F(x)$$

- no access to $\nabla f(x)$ (or to $u \in \partial f(x)$),
- but only to i.i.d. *unbiased* samples, i.e. $\mathbb{E} \nabla F_n(x) = \nabla f(x)$.

Stochastic (sub)gradient descent [Robbins-Monro]

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x) \equiv \mathbb{E} F(x)$$

- no access to $\nabla f(x)$ (or to $u \in \partial f(x)$),
- but only to i.i.d. *unbiased* samples, i.e. $\mathbb{E} \nabla F_n(x) = \nabla f(x)$.

$$x_{n+1} = x_n - \gamma_n \nabla F_n(x_n), \quad \gamma_n = C n^{-\alpha}, \; \alpha \in [0,1] \quad (\textbf{SGD})$$

Stochastic (sub)gradient descent [Robbins-Monro]

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x) \equiv \mathbb{E}F(x)$$

- no access to $\nabla f(x)$ (or to $u \in \partial f(x)$),
- but only to i.i.d. *unbiased* samples, i.e. $\mathbb{E}\nabla F_n(x) = \nabla f(x)$.

$$x_{n+1} = x_n - \gamma_n \nabla F_n(x_n), \quad \gamma_n = Cn^{-\alpha},\ \alpha \in [0,1] \quad (\textbf{SGD})$$

E.g.: Binary classification in Machine Learning

Online learning

$$\nabla F_n(\boldsymbol{\vartheta}) = \nabla F(z_{s(n)}, \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u}_{s(n)} \rangle)$$

for some $s^{(n)} \in \{1, \dots, S\}$.

Stochastic (sub)gradient descent [Robbins-Monro]

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x) \equiv \mathbb{E}F(x)$$

- no access to $\nabla f(x)$ (or to $u \in \partial f(x)$),
- but only to i.i.d. *unbiased* samples, i.e. $\mathbb{E}\nabla F_n(x) = \nabla f(x)$.

$$x_{n+1} = x_n - \gamma_n \nabla F_n(x_n), \quad \gamma_n = Cn^{-\alpha},\ \alpha \in [0,1] \quad (\textbf{SGD})$$

E.g.: Binary classification in Machine Learning

Online learning

$$\nabla F_n(\boldsymbol{\vartheta}) = \nabla F(z_{s^{(n)}}, \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u}_{s^{(n)}} \rangle)$$

for some $s^{(n)} \in \{1, \dots, S\}$.

Batch learning

$$\nabla F_n(\boldsymbol{\vartheta}) = \frac{1}{L} \sum_{\ell=1}^{L} \nabla F(z_{s_\ell^{(n)}}, \langle \boldsymbol{\vartheta}, \phi(\boldsymbol{u}_{s_\ell^{(n)}} \rangle)$$

for L indices $s_\ell^{(n)}$.

Stochastic (sub)gradient descent [Robbins-Monro]

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x) \equiv \mathbb{E} F(x)$$

Theorem (Polyak-Ruppert averaging)

Assume that

- F_n is *convex* and B-Lipschitz continuous on $\{\|x\| \leq D\}$,
- F_n are i.i.d. random functions satisfying $\mathbb{E}\nabla F_n(x) = \nabla f(x)$,
- $\widehat{x}$, the global minimizer, is such that $\|\widehat{x}\| \leq D$.

$$\text{Let} \quad x_{n+1} = P_{\|x\| \leq D}\left(x_n - \frac{2D}{B\sqrt{n}} \nabla F_n(x_n)\right).$$

$$\text{Then for } \overline{x}_{n+1} \equiv \frac{1}{n}\sum_{k=0}^{n} x_k, \quad \mathbb{E} f(\overline{x}_n) - f(\widehat{x}) \leq \frac{2DB}{\sqrt{n}}.$$

Strongly-convex functions

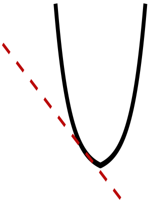
Let $\mathcal{H}$ be a Hilbert space. Let $f\colon \mathcal{H} \to \,]-\infty,+\infty]$ and $\mu > 0$.
f is μ-strongly-convex if

$$(\forall x \in \operatorname{dom} f)(\forall y \in \operatorname{dom} f) \quad f(x) - f(y) \leq \langle \nabla f(x), x - y\rangle - \frac{\mu}{2}\|x-y\|^2.$$

Strongly-convex functions

Let $\mathcal{H}$ be a Hilbert space. Let $f\colon \mathcal{H} \to \left]-\infty, +\infty\right]$ and $\mu > 0$.
f is μ-strongly-convex if

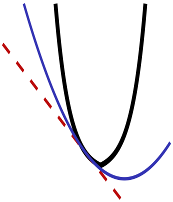
$$(\forall x \in \operatorname{dom} f)(\forall y \in \operatorname{dom} f) \quad f(x) - f(y) \leq \langle \nabla f(x), x - y\rangle - \frac{\mu}{2}\|x-y\|^2.$$



Strongly-convex functions

Let $\mathcal{H}$ be a Hilbert space. Let $f\colon \mathcal{H} \to \,]-\infty,+\infty]$ and $\mu > 0$.
f is μ-strongly-convex if

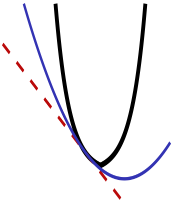
$$(\forall x \in \operatorname{dom} f)(\forall y \in \operatorname{dom} f) \quad f(x) - f(y) \leq \langle \nabla f(x), x - y \rangle - \frac{\mu}{2}\|x - y\|^2.$$



Strongly-convex functions

Let $\mathcal{H}$ be a Hilbert space. Let $f\colon \mathcal{H} \to \,]-\infty, +\infty]$ and $\mu > 0$.
f is μ-strongly-convex if

$$(\forall x \in \operatorname{dom} f)(\forall y \in \operatorname{dom} f) \quad f(x) - f(y) \leq \langle \nabla f(x), x - y \rangle - \frac{\mu}{2}\|x - y\|^2.$$



Characterization of strong-convexity:

Let $f : \mathcal{H} \to]-\infty, +\infty]$ twice differentiable.

- f is μ-strongly-convex iff

  $(\forall x \in \operatorname{dom} f)$, $(\forall z \in \mathcal{H})$

$$\left\langle z \mid \nabla^2 f(x) z \right\rangle \geq \mu \|z\|^2.$$

Stochastic (sub)gradient descent [Robbins-Monro]

$$\widehat{x} \in \underset{x \in \mathcal{H}}{\mathrm{Argmin}}\, f(x) \equiv \mathbb{E}F(x)$$

Theorem (Polyak-Ruppert averaging with strong-convexity)

Assume that

- F_n are i.i.d. random functions satisfying $\mathbb{E}\nabla F_n(x) = \nabla f(x)$,
- F_n is *convex* and B-Lipschitz continuous on $\{\|x\| \leq D\}$,
- f is μ-strongly-convex on $\{\|x\| \leq D\}$,
- $\widehat{x}$, the global minimizer, is such that $\|\widehat{x}\| \leq D$.

$$\text{Let} \quad x_{n+1} = P_{\|x\| \leq D}\left(x_n - \frac{2}{\mu(n+1)} \nabla F_n(x_n)\right).$$

$$\text{Then for } \overline{x}_{n+1} \equiv \frac{2}{n(n+1)} \sum_{k=1}^{n} k x_{k-1}, \quad \mathbb{E}f(\overline{x}_n) - f(\widehat{x}) \leq \frac{2B^2}{\mu(n+1)}.$$

Stochastic forward-backward algorithm

Let $C = \{x \in \mathcal{H} \mid \|x\| \leq D\}$, which is a *convex* subset of $\mathcal{H}$

Stochastic forward-backward algorithm

Let $C = \{x \in \mathcal{H} \mid \|x\| \leq D\}$, which is a *convex* subset of $\mathcal{H}$

Reminder $$\mathrm{prox}_{\iota_C} = P_C$$

Stochastic forward-backward algorithm

Let $C = \{x \in \mathcal{H} \mid \|x\| \leq D\}$, which is a *convex* subset of $\mathcal{H}$

Reminder $$\mathrm{prox}_{\iota_C} = P_C$$

$$x_{n+1} = P_C\left(x_n - \gamma_n \nabla F_n(x_n)\right)$$

Stochastic forward-backward algorithm

Let $C = \{x \in \mathcal{H} \mid \|x\| \leq D\}$, which is a *convex* subset of $\mathcal{H}$

Reminder $\quad \operatorname{prox}_{\iota_C} = P_C$

$$x_{n+1} = P_C\left(x_n - \gamma_n \nabla F_n(x_n)\right) \quad \Leftrightarrow \quad \underbrace{x_{n+1} = \operatorname{prox}_{\gamma_n \iota_C}\left(x_n - \gamma_n \nabla F_n(x_n)\right)}_{\text{forward-backward iteration}}$$

Stochastic forward-backward algorithm

Let $C = \{x \in \mathcal{H} \mid \|x\| \leq D\}$, which is a *convex* subset of $\mathcal{H}$

Reminder $\qquad \operatorname{prox}_{\iota_C} = P_C$

$$x_{n+1} = P_C\left(x_n - \gamma_n \nabla F_n(x_n)\right) \quad \Leftrightarrow \quad \underbrace{x_{n+1} = \operatorname{prox}_{\gamma_n \iota_C}\left(x_n - \gamma_n \nabla F_n(x_n)\right)}_{\text{forward-backward iteration}}$$

General setting: $F_s, g \in \Gamma_0(\mathcal{H})$, F_s differentiable with β-Lipschitz gradient

$$\widehat{x} \in \underset{x \in \mathcal{H}}{\operatorname{Argmin}}\ f(x) = F(x) + g(x) = \frac{1}{S} \sum_{s=1}^{S} F_s(x) + g(x).$$

Stochastic forward-backward algorithm

General setting: $F_s, g \in \Gamma_0(\mathcal{H})$, F_s differentiable with β-Lipschitz gradient

$$\widehat{x} \in \underset{x \in \mathcal{H}}{\operatorname{Argmin}}\ f(x) = F(x) + g(x) = \frac{1}{S} \sum_{s=1}^{S} F_s(x) + g(x).$$

Stochastic forward-backward algorithm

General setting: $F_s, g \in \Gamma_0(\mathcal{H})$, F_s differentiable with β-Lipschitz gradient

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x) = F(x) + g(x) = \frac{1}{S} \sum_{s=1}^{S} F_s(x) + g(x).$$

Stochastic Proximal Gradient (SPG)

Let $\gamma_n = Cn^{-\alpha}$

Stochastic forward-backward algorithm

General setting: $F_s, g \in \Gamma_0(\mathcal{H})$, F_s differentiable with β-Lipschitz gradient

$$\widehat{x} \in \operatorname*{Argmin}_{x \in \mathcal{H}} f(x) = F(x) + g(x) = \frac{1}{S} \sum_{s=1}^{S} F_s(x) + g(x).$$

Stochastic Proximal Gradient (SPG)

Let $\gamma_n = Cn^{-\alpha}$ and $(s^{(n)})_{n \in \mathbb{N}}$ i.i.d. *random* indexes s. t. $\mathbb{P}[s^{(n)} = s] = 1/S$.

Stochastic forward-backward algorithm

General setting: $F_s, g \in \Gamma_0(\mathcal{H})$, F_s differentiable with β-Lipschitz gradient

$$\widehat{x} \in \underset{x \in \mathcal{H}}{\operatorname{Argmin}} \; f(x) = F(x) + g(x) = \frac{1}{S} \sum_{s=1}^{S} F_s(x) + g(x).$$

Stochastic Proximal Gradient (SPG)

Let $\gamma_n = Cn^{-\alpha}$ and $(s^{(n)})_{n \in \mathbb{N}}$ i.i.d. *random* indexes s. t. $\mathbb{P}[s^{(n)} = s] = 1/S$.

$$x_{n+1} = \operatorname{prox}_{\gamma_n g}\left(x_n - \gamma_n \nabla F_{s^{(n)}}(x_n)\right)$$

Stochastic forward-backward algorithm

Stochastic Proximal Gradient (SPG) $F(x) = 1/S \sum_{s=1}^{S} F_s(x)$

Let $\gamma_n = Cn^{-\alpha}$ and $(s^{(n)})_{n\in\mathbb{N}}$ i.i.d. *random* indexes s. t. $\mathbb{P}[s^{(n)} = s] = 1/S$.

$$x_{n+1} = \mathrm{prox}_{\gamma_n g}\left(x_n - \gamma_n \nabla F_{s^{(n)}}(x_n)\right)$$

Theorem

Let $F_s, g \in \Gamma_0(\mathcal{H})$, $\widehat{x} \in \mathrm{Argmin}_{x\in\mathcal{H}} F(x) + g(x)$

Assume that

- F_s differentiable with β-Lipschitz gradient
- F *or* g strongly-convex,
- $\exists\sigma > 0$, $\exists\eta_n > 0$ s.t. $\|\nabla F_s(x) - \nabla F(x)\|^2 \leq \sigma^2 \left(1 + \eta_n \|\nabla F(x)\|^2\right)$,
- $\exists\varepsilon > 0$ s.t. $\gamma_n \leq (1-\varepsilon)\beta^{-1}\left(1 + 2\sigma^2\eta_n\right)^{-1}$.

Then $\mathbb{E}\|x_n - \widehat{x}\|^2 = \mathcal{O}(n^{-\alpha})$ (for $\alpha = 1$, need *well-chosen* C).