Subgradient sampling for nonsmooth and nonconvex minimization

Tam Le (TSE), joint work with Jérôme Bolte (TSE) and Edouard Pauwels (IRIT)

GdR MOA days

1. Gradient method, Stochastic optimization

2. Nonsmooth stochastic optimization

1. Gradient method, Stochastic optimization

2. Nonsmooth stochastic optimization

Nonconvex minimization

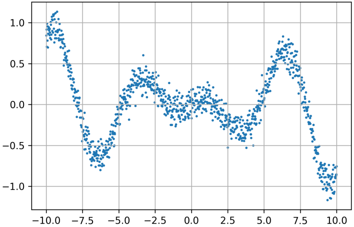
Consider the problem

$$\underset{\mathbf{w}\in\mathbb{R}^p}{\text{Minimize}} \quad F(\mathbf{w}) = \mathbb{E}_{x,y\sim P}\left[\|h(\mathbf{w},x) - y\|_2^2\right]$$

Nonconvex minimization

Consider the problem

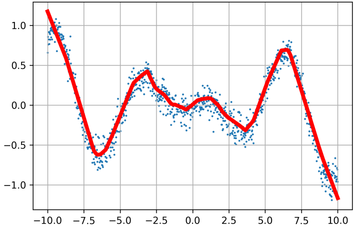
$$\underset{\mathbf{w}\in\mathbb{R}^p}{\text{Minimize}} \quad F(\mathbf{w}) = \mathbb{E}_{x,y\sim P}\left[\|h(\mathbf{w},x) - y\|_2^2\right]$$



Consider the problem

$$\underset{\mathbf{w}\in\mathbb{R}^p}{\text{Minimize}} \quad F(\mathbf{w}) = \mathbb{E}_{x,y\sim P}\left[\| h(\mathbf{w}, x) - y\|_2^2\right]$$

h can be defined several ways: from Linear models to Deep Learning...



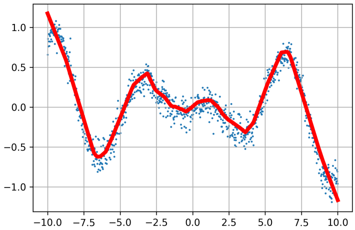
Nonconvex minimization

Consider the problem

$$\underset{\mathbf{w}\in\mathbb{R}^p}{\text{Minimize}} \quad F(\mathbf{w}) = \mathbb{E}_{x,y\sim P}\left[\|h(\mathbf{w},x) - y\|_2^2\right]$$

h can be defined several ways: from Linear models to Deep Learning...

- F is nonconvex
- No access to F in general, only to h and samples from P.



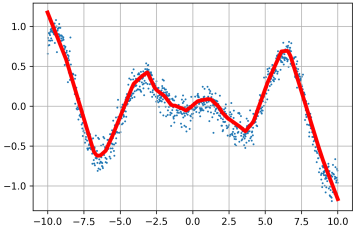
Nonconvex minimization

Consider the problem

$$\underset{\mathbf{w}\in\mathbb{R}^p}{\text{Minimize}} \quad F(\mathbf{w}) = \mathbb{E}_{x,y\sim P}\left[\|h(\mathbf{w},x) - y\|_2^2\right]$$

h can be defined several ways: from Linear models to Deep Learning...

- F is nonconvex
- No access to F in general, only to h and samples from P.



How to find a critical point $\nabla F(w) = 0$?

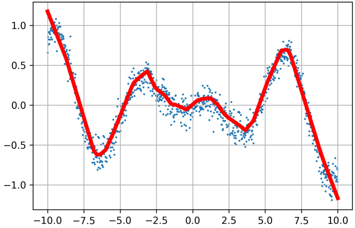
Nonconvex minimization

Consider the problem

$$\underset{\mathbf{w}\in\mathbb{R}^p}{\text{Minimize}} \quad F(\mathbf{w}) = \mathbb{E}_{x,y\sim P}\left[\|h(\mathbf{w},x) - y\|_2^2\right]$$

h can be defined several ways: from Linear models to Deep Learning...

- F is nonconvex
- No access to F in general, only to h and samples from P.



How to find a critical point $\nabla F(w) = 0$?

Classical method: **Gradient method**

$$w_{k+1} = w_k - \alpha_k \nabla F(w_k)$$

Stochastic Gradient descent

But F writes as an expectation,

$$F(w) = \mathbb{E}_{x,y\sim P}\left[\|h(w,x) - y\|_2^2\right] = \mathbb{E}_{\xi\sim P}\left[f(w,\xi)\right]$$

Stochastic Gradient descent

But F writes as an expectation,

$$F(w) = \mathbb{E}_{x,y\sim P}\left[\|h(w,x) - y\|_2^2\right] = \mathbb{E}_{\xi\sim P}\left[f(w,\xi)\right]$$

We do not have access to ∇F !

But F writes as an expectation,

$$F(w) = \mathbb{E}_{x,y\sim P}\left[\|h(w,x) - y\|_2^2\right] = \mathbb{E}_{\xi\sim P}\left[f(w,\xi)\right]$$

We do not have access to ∇F !

Suppose however we can compute f and $\nabla_w f$, and we have i.i.d. samples $(\xi_k)_{k\in\mathbb{N}}$.

Stochastic gradient descent (SGD)

$$w_{k+1} = w_k - \alpha_k \nabla_w f(w_k, \xi_k)$$

But F writes as an expectation,

$$F(w) = \mathbb{E}_{x,y\sim P}\left[\|h(w,x) - y\|_2^2\right] = \mathbb{E}_{\xi\sim P}\left[f(w,\xi)\right]$$

We do not have access to ∇F !

Suppose however we can compute f and $\nabla_w f$, and we have i.i.d. samples $(\xi_k)_{k\in\mathbb{N}}$.

Stochastic gradient descent (SGD)

$$w_{k+1} = w_k - \alpha_k \nabla_w f(w_k, \xi_k)$$

Under reasonable assumptions, we can switch operations $\mathbb{E}$ and ∇, so that $\mathbb{E}_{\xi\sim P}\left[\nabla_w f(w,\xi)\right] = \nabla F(w)$ and

$$\nabla_w f(w_k, \xi_k) \approx \nabla F(w_k)$$

Stochastic Gradient descent

But F writes as an expectation,

$$F(w) = \mathbb{E}_{x,y\sim P}\left[\|h(w,x) - y\|_2^2\right] = \mathbb{E}_{\xi\sim P}\left[f(w,\xi)\right]$$

We do not have access to ∇F !

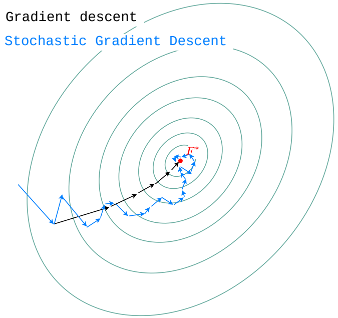
Suppose however we can compute f and $\nabla_w f$, and we have i.i.d. samples $(\xi_k)_{k\in\mathbb{N}}$.

Stochastic gradient descent (SGD)

$$w_{k+1} = w_k - \alpha_k \nabla_w f(w_k, \xi_k)$$

Under reasonable assumptions, we can switch operations $\mathbb{E}$ and ∇, so that $\mathbb{E}_{\xi\sim P}\left[\nabla_w f(w,\xi)\right] = \nabla F(w)$ and

$$\nabla_w f(w_k, \xi_k) \approx \nabla F(w_k)$$



Analysis of the algorithm: the ODE approach

-If $\alpha_k \searrow 0$ but not too quickly, $\sum \alpha_k = +\infty$, $\sum \alpha_k^2 < +\infty$

$$\frac{w_{k+1} - w_k}{\alpha_k} = \underbrace{-\nabla F(w_k) + \varepsilon_k}_{\text{switch } \mathbb{E} \text{ and } \nabla} \qquad \longleftrightarrow \qquad \dot{\gamma} = -\nabla F(\gamma) \tag{1}$$

where ε_k is a "noise" (martingale difference).

-If $\alpha_k \searrow 0$ but not too quickly, $\sum \alpha_k = +\infty$, $\sum \alpha_k^2 < +\infty$

$$\frac{w_{k+1} - w_k}{\alpha_k} = \underbrace{-\nabla F(w_k)}_{\text{switch } \mathbb{E} \text{ and } \nabla} + \varepsilon_k \qquad \longleftrightarrow \qquad \dot{\gamma} = -\nabla F(\gamma) \tag{1}$$

where ε_k is a "noise" (martingale difference).

Meanwhile, along a solution γ of (1)

$$\frac{d}{dt}(F \circ \gamma)(t) \underset{\text{chain-rule}}{=} \langle \nabla F(\gamma(t)), \dot{\gamma}(t) \rangle = -\|\nabla F(\gamma(t))\|^2 \leq 0. \tag{2}$$

Analysis of the algorithm: the ODE approach

-If $\alpha_k \searrow 0$ but not too quickly, $\sum \alpha_k = +\infty$, $\sum \alpha_k^2 < +\infty$

$$\frac{w_{k+1} - w_k}{\alpha_k} = \underbrace{-\nabla F(w_k)}_{\text{switch } \mathbb{E} \text{ and } \nabla} + \varepsilon_k \qquad \longleftrightarrow \qquad \dot{\gamma} = -\nabla F(\gamma) \tag{1}$$

where ε_k is a “noise” (martingale difference).

Meanwhile, along a solution γ of (1)

$$\frac{d}{dt}(F \circ \gamma)(t) \underset{\text{chain-rule}}{=} \langle \nabla F(\gamma(t)), \dot{\gamma}(t) \rangle = -\|\nabla F(\gamma(t))\|^2 \leq 0. \tag{2}$$

Suppose (w_k) is bounded a.s.

Analysis of the algorithm: the ODE approach

-If $\alpha_k \searrow 0$ but not too quickly, $\sum \alpha_k = +\infty$, $\sum \alpha_k^2 < +\infty$

$$\frac{w_{k+1} - w_k}{\alpha_k} = \underbrace{-\nabla F(w_k) + \varepsilon_k}_{\text{switch } \mathbb{E} \text{ and } \nabla} \qquad \longleftrightarrow \qquad \dot{\gamma} = -\nabla F(\gamma) \tag{1}$$

where ε_k is a "noise" (martingale difference).

Meanwhile, along a solution γ of (1)

$$\frac{d}{dt}(F \circ \gamma)(t) \underset{\text{chain-rule}}{=} \langle \nabla F(\gamma(t)), \dot{\gamma}(t) \rangle = -\|\nabla F(\gamma(t))\|^2 \leq 0. \tag{2}$$

Suppose (w_k) is bounded a.s.

- **Convergence to the critical set.** Combining (1) and (2), accumulation points of $(w_k)_{k \in \mathbb{N}}$ is a connected set and are critical, $(\nabla F(w) = 0)$.

Analysis of the algorithm: the ODE approach

-If $\alpha_k \searrow 0$ but not too quickly, $\sum \alpha_k = +\infty$, $\sum \alpha_k^2 < +\infty$

$$\frac{w_{k+1} - w_k}{\alpha_k} = \underbrace{-\nabla F(w_k) + \varepsilon_k}_{\text{switch } \mathbb{E} \text{ and } \nabla} \qquad \longleftrightarrow \qquad \dot{\gamma} = -\nabla F(\gamma) \tag{1}$$

where ε_k is a "noise" (martingale difference).

Meanwhile, along a solution γ of (1)

$$\frac{d}{dt}(F \circ \gamma)(t) \underset{\text{chain-rule}}{=} \langle \nabla F(\gamma(t)), \dot{\gamma}(t) \rangle = -\|\nabla F(\gamma(t))\|^2 \leq 0. \tag{2}$$

Suppose (w_k) is bounded a.s.

- **Convergence to the critical set.** Combining (1) and (2), accumulation points of $(w_k)_{k \in \mathbb{N}}$ is a connected set and are critical, $(\nabla F(w) = 0)$.

- **Convergence of the objective function.** Suppose furthermore the set of critical values ($F(w)$ such that $\nabla F(w) = 0$) has empty interior (Sard's condition), then $F(w_k)$ converges (empty interior + connected).

1 Gradient method, Stochastic optimization

2 Nonsmooth stochastic optimization

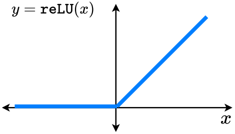
Why should we care about nonsmooth functions?

In deep learning, predictions are built upon compositions.

$$h(w,x) = \sigma(A_L\sigma(A_{L-1}\ldots\sigma(A_2\sigma(A_1x+b_1)+b_2)+b_{L-1}\ldots)+b_L)$$

$w = (A_1, A_2 \ldots A_L, b_1, \ldots, b_L)$. σ are **nonsmooth** because defined with conditional statements.

$$\texttt{reLU}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

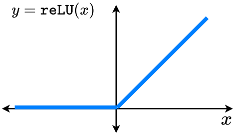


In deep learning, predictions are built upon compositions.

$$h(w,x) = \sigma(A_L\sigma(A_{L-1}\ldots\sigma(A_2\sigma(A_1x+b_1)+b_2)+b_{L-1}\ldots)+b_L)$$

$w = (A_1, A_2 \ldots A_L, b_1, \ldots, b_L)$. σ are **nonsmooth** because defined with conditional statements.

$$\mathtt{reLU}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$



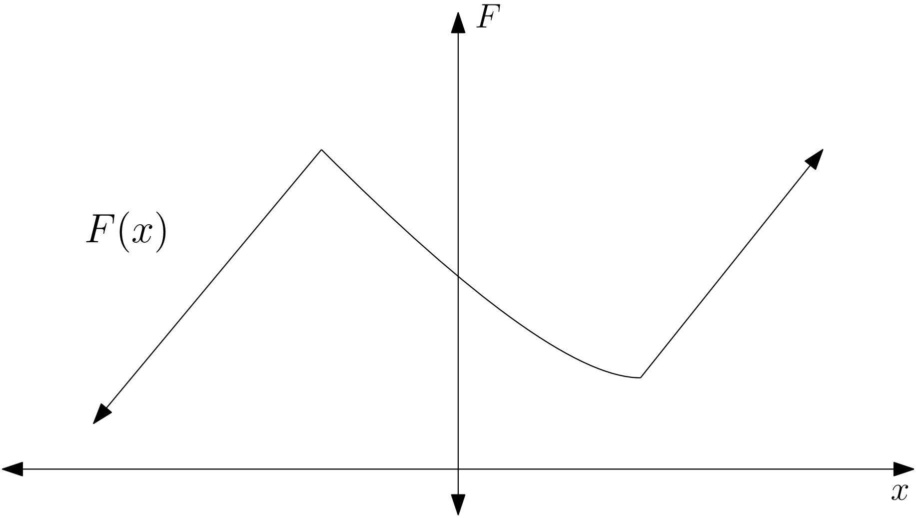
In this setting, can we have some kind of stochastic gradient method:

$$w_{k+1} = w_k - \alpha v(w_k, \xi_k)$$

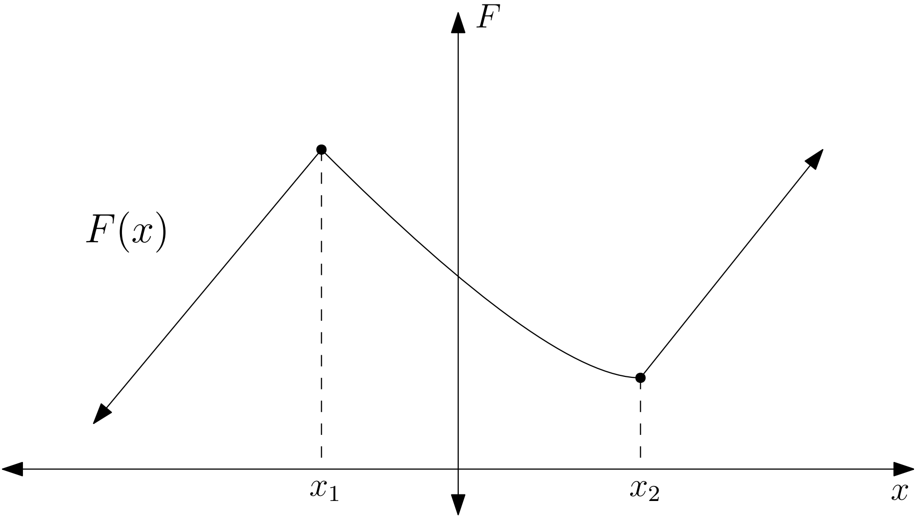
where $v(w_k, \xi_k)$ approximates a **gradient-like** object for $F(w_k)$?

The Clarke subgradient

How to define a gradient-like object where F is non-differentiable ?

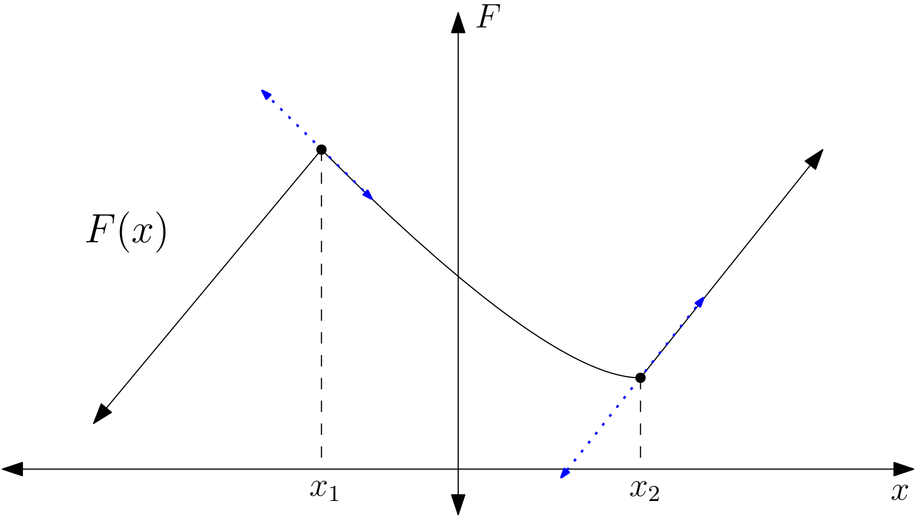


How to define a gradient-like object where F is non-differentiable ?



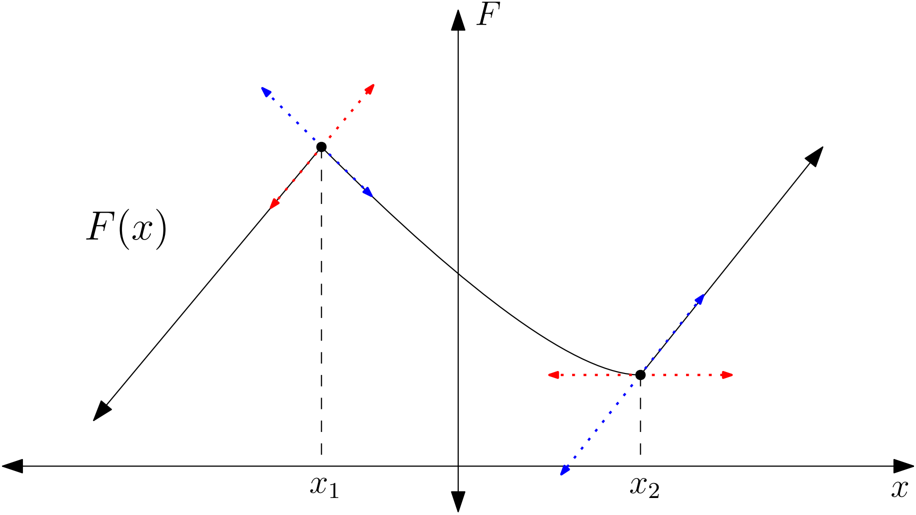
The Clarke subgradient

How to define a gradient-like object where F is non-differentiable ?

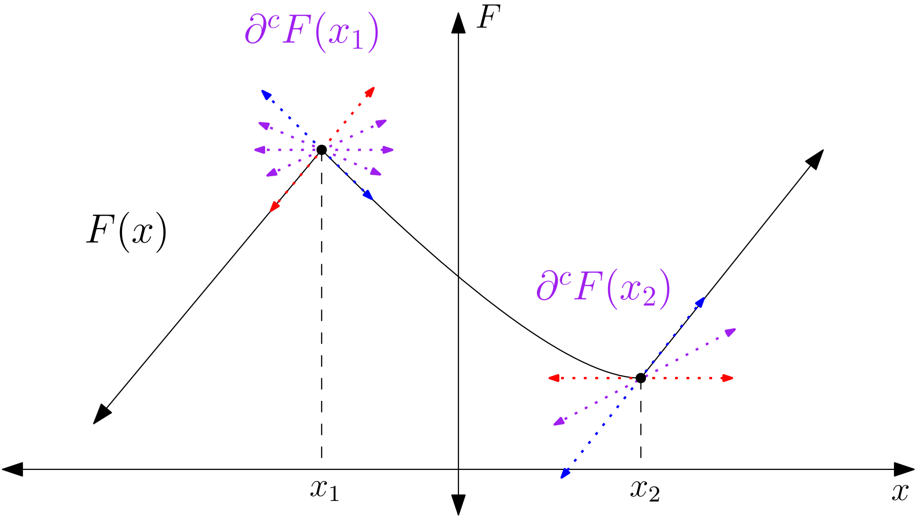


The Clarke subgradient

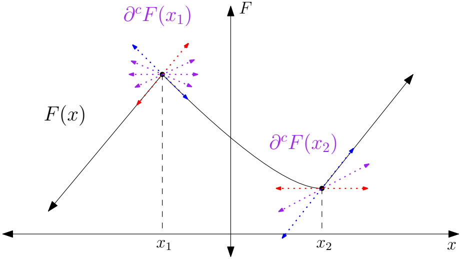
How to define a gradient-like object where F is non-differentiable ?



How to define a gradient-like object where F is non-differentiable ?



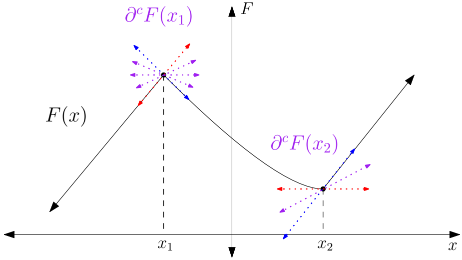
How to define a gradient-like object where F is non-differentiable ?



Let $F : \mathbb{R}^n \to \mathbb{R}$ Lipschitz, differentiable on diff_F of full measure. The **Clarke subgradient** is

$$\partial^c F(x) = \mathrm{conv}\left\{ \lim_{k \to +\infty} \nabla F(x_k) : x_k \in \mathrm{diff}_F, x_k \underset{k \to +\infty}{\to} x \right\}$$

How to define a gradient-like object where F is non-differentiable ?



Let $F : \mathbb{R}^n \to \mathbb{R}$ Lipschitz, differentiable on diff_F of full measure. The **Clarke subgradient** is

$$\partial^c F(x) = \mathrm{conv}\left\{ \lim_{k\to+\infty} \nabla F(x_k) : x_k \in \mathrm{diff}_F, x_k \underset{k\to+\infty}{\to} x \right\}$$

we "fill the holes".

Clarke subgradients are "set-valued"

Clarke subgradients are “set-valued”

$+$ **operation is set-valued**

Set-valued gradients imply a set-valued formalism

Clarke subgradients are "set-valued"

- **operation is set-valued**

$$\partial^c F + \partial^c G := \{A + B \mid A \in \partial^c F, B \in \partial^c G\}$$

Clarke subgradients are "set-valued"

$+$ **operation is set-valued**

$$\partial^c F + \partial^c G := \{A + B \mid A \in \partial^c F, B \in \partial^c G\}$$

Integrals, expectations are set-valued

$$\mathbb{E}_{\xi \sim P}[\partial^c F(\xi)] = \left\{ \int_{\mathbb{R}^m} v(s) dP(s) : v(s) \in \partial^c F(s), v \text{ integrable} \right\}$$

Clarke subgradients are "set-valued"

$+$ **operation is set-valued**

$$\partial^c F + \partial^c G := \{A + B \mid A \in \partial^c F, B \in \partial^c G\}$$

Integrals, expectations are set-valued

$$\mathbb{E}_{\xi \sim P}[\partial^c F(\xi)] = \left\{ \int_{\mathbb{R}^m} v(s) dP(s) : v(s) \in \partial^c F(s), v \text{ integrable} \right\}$$

First-order optimality condition

Smooth: $0 = \nabla F(x)$, Nonsmooth: $0 \in \partial^c F(x)$

Clarke subgradients are "set-valued"

$+$ **operation is set-valued**

$$\partial^c F + \partial^c G := \{A + B \mid A \in \partial^c F, B \in \partial^c G\}$$

Integrals, expectations are set-valued

$$\mathbb{E}_{\xi \sim P}[\partial^c F(\xi)] = \left\{ \int_{\mathbb{R}^m} v(s) dP(s) : v(s) \in \partial^c F(s), v \text{ integrable} \right\}$$

First-order optimality condition

Smooth: $0 = \nabla F(x)$, Nonsmooth: $0 \in \partial^c F(x)$

Differential equations become inclusions

Smooth: $\dot{x}(t) = -\nabla F(x(t))$, Nonsmooth: $\dot{x}(t) \in -\partial^c F(x(t))$ a.e. in t

Can we use the subgradient formally ?

$$w_{k+1} \in w_k - \alpha_k \partial^c_w f(w_k, \xi_k)$$

Failure of the Clarke subgradient in stochastic optimization

Can we use the subgradient formally ?

$$w_{k+1} \in w_k - \alpha_k \partial^c_w f(w_k, \xi_k)$$

No !!!

If $\partial^c_w f(w, \xi)$ is the subgradient of f with respect to w, then

$$\partial^c F(w) \subset \mathbb{E}_{\xi \sim P}\left[\partial^c_w f(w, \xi)\right]$$

$\rightarrow$ **Subgradient sampling is not consistent.**

Can we use the subgradient formally ?

$$w_{k+1} \in w_k - \alpha_k \partial_w^c f(w_k, \xi_k)$$

No !!!

If $\partial_w^c f(w, \xi)$ is the subgradient of f with respect to w, then

$$\partial^c F(w) \subset \mathbb{E}_{\xi \sim P} [\partial_w^c f(w, \xi)]$$

$\rightarrow$ **Subgradient sampling is not consistent.**

There exist (a lot of) functions F, differentiable almost everywhere, such that $\partial^c F = [-1, 1]$ everywhere (see Borwein and Wang 2000).

$\rightarrow$ **Sometimes, the subgradient does not contain variational information**

Failure of the Clarke subgradient in stochastic optimization

Can we use the subgradient formally ?

$$w_{k+1} \in w_k - \alpha_k \partial^c_w f(w_k, \xi_k)$$

No !!!

If $\partial^c_w f(w, \xi)$ is the subgradient of f with respect to w, then

$$\partial^c F(w) \subset \mathbb{E}_{\xi \sim P}\left[\partial^c_w f(w, \xi)\right]$$

$\rightarrow$ **Subgradient sampling is not consistent.**

There exist (a lot of) functions F, differentiable almost everywhere, such that $\partial^c F = [-1, 1]$ everywhere (see Borwein and Wang 2000).

$\rightarrow$ **Sometimes, the subgradient does not contain variational information**

We need a new notion of gradient, which comes along regularity.

Let $F : \mathbb{R}^n \to \mathbb{R}$ locally Lipschitz, and a set-valued map $D_F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$. D_F is a **conservative gradient** for F if

Conservative gradients: definition

Let $F : \mathbb{R}^n \to \mathbb{R}$ locally Lipschitz, and a set-valued map $D_F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$. D_F is a **conservative gradient** for F if

- D_F is graph closed, nonempty valued
- D_F is locally bounded
- (D_F convex)

$\to$ Existence of solutions for $\dot{x} \in -D_F(x)$

Let $F : \mathbb{R}^n \to \mathbb{R}$ locally Lipschitz, and a set-valued map $D_F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$. D_F is a **conservative gradient** for F if

- D_F is graph closed, nonempty valued
- D_F is locally bounded
- (D_F convex)

$\to$ Existence of solutions for $\dot{x} \in -D_F(x)$

Chain rule along curves

For all absolutely continuous curve $\gamma : [0, 1] \to \mathbb{R}^n$,

$$\frac{\mathrm{d}}{\mathrm{d}t}(F \circ \gamma)(t) = \langle v, \dot{\gamma}(t)\rangle \text{ for all } v \in D_F(\gamma(t)),$$

for almost all $t \in [0, 1]$.

Conservative gradients: definition

Let $F : \mathbb{R}^n \to \mathbb{R}$ locally Lipschitz, and a set-valued map $D_F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$. D_F is a **conservative gradient** for F if

- D_F is graph closed, nonempty valued
- D_F is locally bounded
- (D_F convex)

$\to$ Existence of solutions for $\dot{x} \in -D_F(x)$

Chain rule along curves
For all absolutely continuous curve $\gamma : [0,1] \to \mathbb{R}^n$,

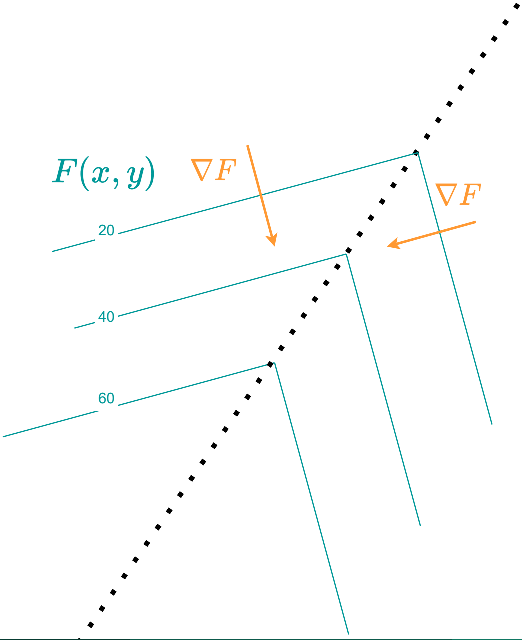
$$\frac{\mathrm{d}}{\mathrm{d}t}(F \circ \gamma)(t) = \langle v, \dot{\gamma}(t) \rangle \text{ for all } v \in D_F(\gamma(t)),$$

for almost all $t \in [0,1]$.

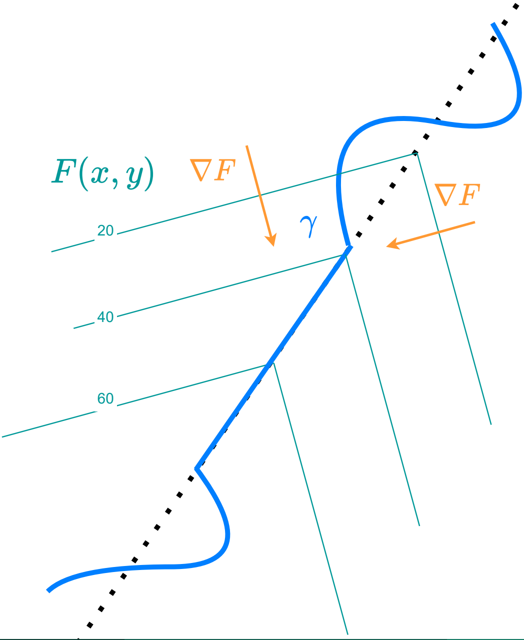
- F is called **path differentiable**
- If F is path differentiable, then $\partial^c F$ **is a conservative gradient**.
- D_F is not unique !

Can we have sufficient conditions for this chain rule?

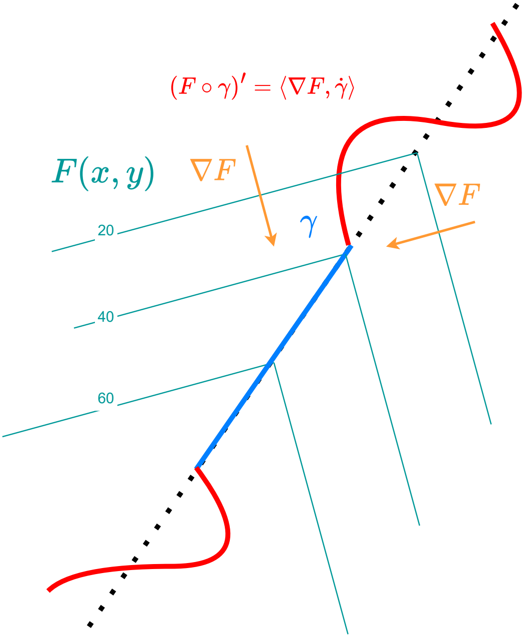
Path differentiability of piecewise affine functions



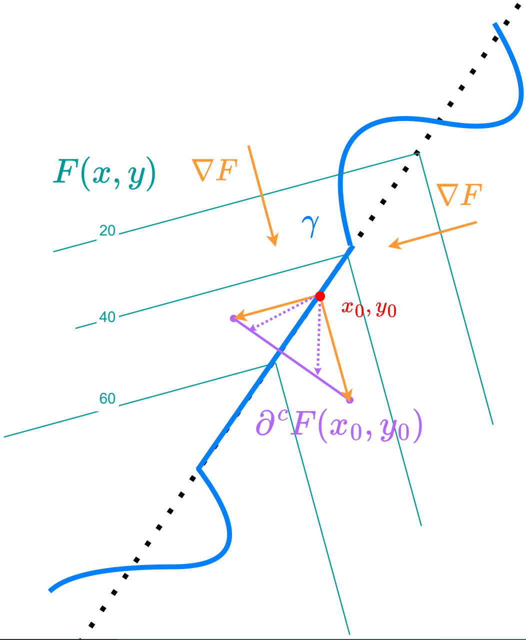
Path differentiability of piecewise affine functions



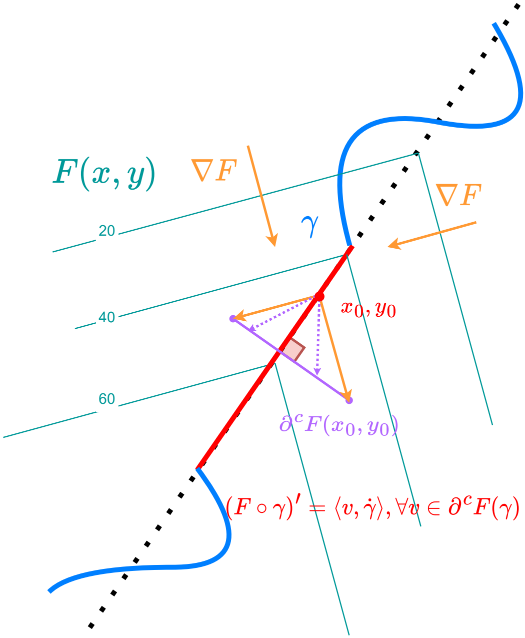
Path differentiability of piecewise affine functions



Path differentiability of piecewise affine functions

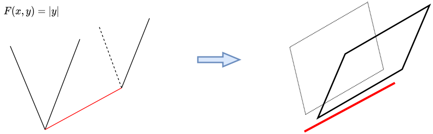


Path differentiability of piecewise affine functions



Path differentiability in practice

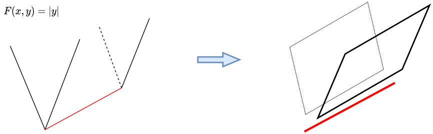
Graphs of piecewise affine functions can be divided into affine pieces.



They are furthermore built with simple operations $(\leq, =, -, +, \cdot, \text{if}, \text{else})$, see for instance reLU function.

Path differentiability in practice

Graphs of piecewise affine functions can be divided into affine pieces.



They are furthermore built with simple operations $(\leq, =, -, +, \cdot, \text{if}, \text{else})$, see for instance reLU function.

We guess some relation

"Compositional formula $\sim$ Graph structure"

Definable geometry generalizes this equivalence for other dictionary of operations. For instance, suppose we have a function implemented with

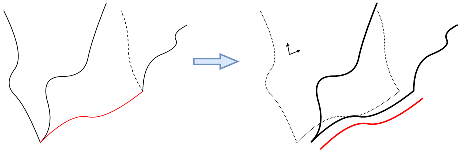
$$(\mathrm{exp}, \mathrm{log}, \leq, =, -, +, \cdot, \times, \mathrm{if}, \mathrm{else})$$

we call it a definable function.

Definable geometry generalizes this equivalence for other dictionary of operations. For instance, suppose we have a function implemented with

$$(\exp, \log, \leq, =, -, +, \cdot, \times, \text{if}, \text{else})$$

we call it a definable function. Then by **Definable geometry**, its graph can be divided into nice pieces (C^r-manifolds) and locally looks like piecewise affine functions.



Definable functions are path differentiable.

Compatibility with subgradient sampling

Theorem (Interchanging $\mathbb{E}$ and conservative gradient)

Suppose f definable. Under measurability, integrability assumptions,

$\mathbb{E}_{\xi\sim P}\left[\partial^c_w f(\cdot,\xi)\right]$ *is a* **conservative gradient** *for* $F := \mathbb{E}_{\xi\sim P}[f(\cdot,\xi)]$.

Sampling $\partial^c_w f(w,\xi)$, $\xi \sim P$ averages a descent direction at w

Application to stochastic optimization

We study the convergence of **nonsmooth SGD:**

$$w_{k+1} \in w_k - \alpha_k \partial_w^c f(w_k, \xi_k)$$

We study the convergence of **nonsmooth SGD:**

$$w_{k+1} \in w_k - \alpha_k \partial_w^c f(w_k, \xi_k)$$

Main assumptions

- $\sum \alpha_k = +\infty$, $\alpha_k \to 0$
- $f : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$ is definable *(semialgebraic, globally subanalytic)*.
- Integrability assumption: For almost all $s \in \mathbb{R}^m$, $x, y \in \mathbb{R}^p$

$$|f(x,s) - f(y,s)| \leq \kappa(s)(1 + (\|x\| + \|y\|)^r)\|x - y\|$$

κ^n is P-integrable for all $n \in \mathbb{N}$.

Theorem (Weak convergence of nonsmooth stochastic gradient descent)

Let (w_k) generated by nonsmooth SGD. Suppose (w_k) is bounded a.s., then any essential accumulation point $\mathbf{a}$ of (w_k) satisfies

$$0 \in \mathbb{E}_{\xi \sim P} \left[\partial_w^c f(\mathbf{a}, \xi)\right], \ a.s.$$

a is an essential accumulation point if for all neighborhood U of **a**,

$$\limsup_{k\to\infty} \frac{\sum_{i=0}^{k} \alpha_i \mathbf{1}_{w_i \in U}}{\sum_{i=0}^{k} \alpha_i} > 0$$

interpretation: the proportion of time spent around **a** doesn't vanish as $k \to \infty$.

a is an essential accumulation point if for all neighborhood U of **a**,

$$\limsup_{k\to\infty} \frac{\sum_{i=0}^{k} \alpha_i \mathbf{1}_{w_i \in U}}{\sum_{i=0}^{k} \alpha_i} > 0$$

interpretation: the proportion of time spent around **a** doesn't vanish as $k \to \infty$.

Suppose furthermore

- $\sum \alpha_i^2 < +\infty$
- $P \ll \lambda$ has a definable density, with compact support $\to$ Sard's condition (*"Definability" of integrals, Cluckers and Miller 2009*).

a is an essential accumulation point if for all neighborhood U of **a**,

$$\limsup_{k\to\infty} \frac{\sum_{i=0}^{k} \alpha_i \mathbf{1}_{w_i \in U}}{\sum_{i=0}^{k} \alpha_i} > 0$$

interpretation: the proportion of time spent around **a** doesn't vanish as $k \to \infty$.

Suppose furthermore

- $\sum \alpha_i^2 < +\infty$
- $P \ll \lambda$ has a definable density, with compact support $\rightarrow$ Sard's condition (*"Definability" of integrals, Cluckers and Miller 2009*).

Theorem (Convergence of stochastic subgradient descent)

*$F(w_k)$ converges, and any accumulation point **a** of (w_k) satisfies*

$$0 \in \mathbb{E}_{\xi\sim P}\left[\partial^c f(\mathbf{a}, \xi)\right], \textit{a.s.}$$

a is an essential accumulation point if for all neighborhood U of **a**,

$$\limsup_{k\to\infty} \frac{\sum_{i=0}^{k} \alpha_i \mathbf{1}_{w_i \in U}}{\sum_{i=0}^{k} \alpha_i} > 0$$

interpretation: the proportion of time spent around **a** doesn't vanish as $k \to \infty$.

Suppose furthermore

- $\sum \alpha_i^2 < +\infty$
- $P \ll \lambda$ has a definable density, with compact support $\to$ Sard's condition (*"Definability" of integrals, Cluckers and Miller 2009*).

Theorem (Convergence of stochastic subgradient descent)

$F(w_k)$ converges, and any accumulation point **a** *of (w_k) satisfies*

$$0 \in \mathbb{E}_{\xi\sim P}\left[\partial^c f(\mathbf{a},\xi)\right], a.s.$$

Genericity of Clarke criticality

Question: If F has conservative gradient D_F, then $D_F = \nabla F$ almost everywhere. Can we have $0 \in \partial^c F(\mathbf{a})$ instead?
Answer: Randomizing w_0, (α_k) is sufficient.

Theorem (Genericity of Clarke criticality)

Suppose $\alpha_k = \frac{\alpha_0}{k+1}$. When randomizing w_0 and α_0 (Gaussian or uniformly) then a.s., any accumulation point $\mathbf{a}$ satisfies

$$0 \in \partial^c F(\mathbf{a})$$

Conclusion

- The theory of conservative gradients allows to study stochastic subgradient methods on nonsmooth functions since it encompasses its keystone principles:

- descent mechanism (chain rule along curves)
- compatibility with (sub)gradient sampling

Conclusion

- The theory of conservative gradients allows to study stochastic subgradient methods on nonsmooth functions since it encompasses its keystone principles:

- descent mechanism (chain rule along curves)
- compatibility with (sub)gradient sampling

- Definable theory allows to retrieve classical subgradient criticality with randomized initialization.

Conclusion

- The theory of conservative gradients allows to study stochastic subgradient methods on nonsmooth functions since it encompasses its keystone principles:

  - descent mechanism (chain rule along curves)
  - compatibility with (sub)gradient sampling

- Definable theory allows to retrieve classical subgradient criticality with randomized initialization.
- Gradient descent is widely used in Deep Learning with nonsmooth gradient oracle like backpropagation or implicit differentiation $\rightarrow$ conservative gradient also models these.

Subgradient sampling for Nonsmooth Nonconvex minimization
`https://arxiv.org/abs/2202.13744`

Thanks for listening!