Descente de gradient généralisée avec un coût de transport optimal.

Joël GARDE Olivier FERCOQ
LTCI, Télécom Paris, 91120 Palaiseau, France

Résumé – Cet article présente un algorithme qui généralise la descente de gradient sur les mesures de probabilité. La mise en œuvre de cet algorithme à pour objectif de résoudre des problèmes inverses en localisation de sources ou en apprentissage de dictionnaire. Ces problèmes sont résolus hors-grille, c'est-à-dire sans discrétisation de l'espace des sources, opérant à la place avec des mesures discrètes sur ce même espace. L'algorithme présenté s'apparente à une version discrète en temps d'un flot de gradient de Wasserstein, et on l'emploiera pour la résolution d'un exemple de problème de localisation de sources.

Abstract – We present a generalization of gradient descent applied to probability measures. This algorithm is used to solve the source localization inverse problem or in the dictionnary learning problem. Both problems are solved in a grid-free manner, meaning without discretizing the source space, opting instead to use discrete measures supported on the source space. The algorithm we propose relates to a discretized version of a Wasserstein gradient flow, and we use it to solve a toy source localisation problem.

1 Introduction

En localisation de source comme en apprentissage de dictionnaire, on s'intéresse à la décomposition d'un signal $y \in \mathbf{R}^m$ comme la somme

$$y = \sum_{i=1}^{n} \phi(x_i) \tag{1}$$

d'éléments $\phi(x_1), \dots, \phi(x_n)$ d'un dictionnaire $\phi\colon X \to \mathbf{R}^m$. En posant la mesure discrète $\alpha = \sum_{i=1}^n \delta(x_i)$ pour la mesure de Dirac δ, retrouver les sources $x_1, \dots, x_n$, c'est minimiser la fonction convexe $F(\alpha) = \|\int \phi \,\mathrm{d}\alpha - y\|^2$ sur les mesures de probabilité $\alpha \in P(\mathbf{R}^m)$. Il nous faut donc un algorithme pour minimiser une fonction convexe sur $P(\mathbf{R}^m)$. Entre autre, c'est l'algorithme de Frank-Wolfe [2, 4] ou une discrétisation d'un flot de gradient de Wasserstein [3] qui sont utilisés en régression parcimonieuse. Ce dernier admet une implémentation particulièrement simple comme descente de gradient non convexe sur les positions (et poids, ici omis), de particules.

Une descente de gradient miroir correspondrait à quel type de flot de Wasserstein ? Si l'implémentation est simple, la description peut-elle l'être tout autant ? On se propose d'étudier un algorithme en temps discret, court-circuitant la nécessité de définir le gradient de Wasserstein et son flot.

Contributions. Dans cet article, on donne une description des fonctions sur les mesures qui sont concaves pour le coût de transport, propriété qui généralise les fonctions lisses. On obtient alors une implémentation de l'algorithme de descente de gradient généralisée [8] pour les fonctions convexes sur les mesures. Cet algorithme est finalement testé sur un problème de localisation de sources.

2 Transport et concavité généralisée

Cette section établit les résultats de base concernant le transport optimal et la c-concavité, sur lesquels reposent nos développements.

Notations. Pour une fonction $f\colon X \to \overline{\mathbf{R}}$, on appelle sa conjuguée convexe la fonction f^* définie par $f^*(y) = \sup_X \langle x, y\rangle - f(x)$. Pour une mesure μ, sa mesure image $f\#\mu$ est donnée par $\int g \,\mathrm{d}(f\#\mu) = \int (g \circ f)\,\mathrm{d}\mu$ pour toute fonction test g continue.

Coût de transport. On se dote d'une fonction de couplage $c\colon X \times Y \to [0, +\infty[$ continue et positive définie sur deux espaces polonais X et Y, et de deux mesures $\alpha \in P(X)$, $\beta \in P(Y)$. Un *plan de transport* $\pi \in \Pi(\alpha, \beta)$ est une mesure de probabilité sur $X \times Y$ avec pour marginales α et β : il décrit comment transporter α sur β. Le *coût de transport* [11]

$$\mathcal{K}_c(\alpha, \beta) = \inf \int_{X\times Y} c(x, y)\,\mathrm{d}\pi(x, y) : \pi \in \Pi(\alpha, \beta) \tag{2}$$

admet une formulation duale

$$\mathcal{T}_c(\alpha, \beta) = \sup_{f,g} \int_X f \,\mathrm{d}\alpha + \int_{X^*} g \,\mathrm{d}\beta \tag{3}$$

où les *potentiels* f et g obéissent $f(x) + g(y) \leq c(x, y)$. On peut restreindre le choix aux potentiels *conjugués*

$$\begin{cases} f(x) = g^c(x) = \inf_{y\in Y} c(x, y) - g(y) \\ g(y) = f^c(y) = \inf_{x\in X} c(x, y) - f(x). \end{cases} \tag{4}$$

Si c'est le cas, on en déduit que $f^{cc} = f$, et cette dernière équation définit l'ensemble $\Gamma_c(X)$ des fonctions *c-concaves*.

Fonctions conjuguées. La conjuguée en c [9], $f \mapsto f^c$ définie comme en (4) par

$$f^c(y) = \inf_{x\in X} c(x, y) - f(x) \tag{5}$$

généralise celle de Fenchel, donc elle a des propriétés analogues. Notamment, on définit le c-surdifférentiel $\partial_c f$ de $f\colon X \to \mathbf{R}$ comme l'ensemble des $(x, y) \in X \times Y$ tels que

$$f(x) + f^c(y) = c(x, y). \tag{6}$$

théorème 1.

$$y \in \partial_c f(x) \iff x \in \operatorname{argmin}_{x' \in X} \{c(x', y) - f(x')\} \quad (7)$$

$$\partial_c f^c = (\partial_c f)^{-1} \iff f^{cc} = f. \quad (8)$$

Démonstration. Il existe $y \in \partial_c f(x)$ si et seulement si $f^c(y) = \inf_{x' \in X} c(x', y) - f(x') = c(x, y) - f(x)$. Ensuite, on utilisera $f^{cc} \geq f$ et $f(x) + f^c(y) \leq c(x, y)$ [11, 1.34] : alors $f(x) + f^c(y) = c(x, y)$ implique $f^{cc}(x) + f^c(y) = c(x, y)$ et si $f = f^{cc}$ la réciproque est également vraie. □

théorème 2. *Soit $x \in X$, alors*

$$\partial_c f(x) \neq \emptyset \implies f^{cc}(x) = f(x) \quad (9)$$

et si la fonction $y \mapsto c(x, y)$ est continue et Y est compact alors la réciproque est vraie,

$$\partial_c f(x) \neq \emptyset \iff f^{cc}(x) = f(x). \quad (10)$$

Démonstration. Soit $y \in \partial_c f(x)$. Alors $f(x) + f^c(y) = c(x, y)$ et $f^{cc}(x) \geq f(x)$, donc $f^{cc}(x) + f^c(y) = c(x, y)$ Pour la réciproque, soit f tel que $f^{cc} = f$. Par 8 puis (7) il vient $\partial_c f(x) = \operatorname{argmin}_Y c(x, y) - f^c(y)$. Quand $y \mapsto c(x, y)$ est continue, (donc semi-continue supérieure) alors $f^c(y) = \inf_x c(x, y) - f(x)$ l'est aussi : $y \mapsto c(x, y) - f^c(y)$ est semi-continue inférieurement sur un compact, donc atteint un minimum. □

Conjuguées pour l'écart de Fenchel. On décrit ici une famille de fonctions de couplage pour lesquelles la conjugaison peut être calculée explicitement Soit $h \colon X \to \mathbf{R}$. On définit *L'écart de Fenchel* pour h par

$$\begin{aligned} \mathcal{F}(h) \colon X \times X^* &\to [0, +\infty] \\ (x, y) &\mapsto h(x) + h^*(y) - \langle x, y \rangle. \end{aligned} \quad (11)$$

Supposons que h soit une fonction *Legendre* (l'essentiel est que son gradient ∇h réalise une bijection entre $X = \operatorname{int}(\operatorname{dom} h)$ et $X^* = \operatorname{int}(\operatorname{dom} h^*)$, d'inverse ∇h^*). Alors, $\mathcal{F}(h)(x, \nabla h(z)) = h(x) - h(z) - \langle \nabla h(z), x - z \rangle$ est une *divergence de Bregman*. Les plus connues sont le coût quadratique $h = (1/2) \|\cdot\|^2$ pour laquelle $\mathcal{F}(h)(x, y) = (1/2) \|x - y\|^2$ et l'entropie négative $h(p) = -\sum_i p_i \log p_i$ pour laquelle $\mathcal{F}(h)(p, q) = \mathrm{KL}(p, \operatorname{softmax}(q))$ est la divergence de Kullback-Leibler ; L'article [1] détaille ces résultats.

Notons $f^h(y) = \inf_{x \in X} \mathcal{F}(h)(x, y) - f(x)$ la $\mathcal{F}(h)$-conjugée de $f \colon X \to \mathbf{R}$.

théorème 3.

$$f^h(y) = h^*(y) - (h - f)^*(y) \quad (12)$$

$$f^{hh} = f \iff (h - f)^{**} = (h - f) \quad (13)$$

$$\partial_h f = \partial(h - f) \quad (14)$$

Démonstration. On calcule $f^h(y) = \inf_{x \in X} \{\mathcal{F}(h)(x, y) - f(x)\} = h^*(y) + \inf_{x \in X} \{h(x) - \langle x, y \rangle - f(x)\} = h^*(y) - (h - f)^*(y)$, et de même, $f^{hh} = h - (h - f)^{**}$. Développer $f(x) + f^h(y) = \mathcal{F}(h)(x, y)$ en $(h - f)(x) + (h - f)^*(y) = \langle x, y \rangle$ révèle l'expression du surdifférentiel. □

L'équation (13) dit que f est $\mathcal{F}(h)$-concave si et seulement si f est h-lisse, ce qui généralise [8, 4.5] aux fonctions non différentiables. Nous avons désormais une description de la c-concavité suffisante précise pour résoudre les problèmes de transport avec un coût $\mathcal{F}(h)$.

théorème 4. *Soit $h \colon X \to \overline{\mathbf{R}}$ telle que $h^{**} = h$. On se restreint à $\operatorname{int} \operatorname{dom} \mathcal{F}(h) \subset X \times Y$ de sorte que le coût $\mathcal{F}(h)$ soit inférieurement semi-continu et à valeurs finies. Un plan $\pi \in \Pi(\mu, \nu)$ est optimal si et seulement s'il est concentré sur le sous-différentiel d'une fonction convexe φ. De plus, il existe un plan optimal π et un potentiel c-concave f optimal tels que les égalités suivantes soient satisfaites*

$$\int c \, \mathrm{d}\pi = \int f \, \mathrm{d}\mu + \int f^h \, \mathrm{d}\nu \quad (15)$$

$$\operatorname{conc}(\pi) \subset \partial\varphi \quad (16)$$

$$f = h - \varphi, \quad f^h = h^* - \varphi^* \quad (17)$$

Démonstration. Ce théorème est l'application de [11, 1.6.2], et [12] au coût $\mathcal{F}(h)$ grâce au théorème 3. □

3 Descente de gradient généralisée

Une fonction c-concave est l'enveloppe supérieure d'une famille de *fonctions élémentaires* de la forme $x \mapsto c(x, y) + a$, $a \in \mathbf{R}$ [9]. On peut donc employer une stratégie de Majoration-Minimisation qui sélectionne la fonction élémentaire tangente (Majoration) et la minimise dans un second temps (Minimisation). Cette Majoration-Minimisation correspond à la *descente de gradient avec coût généralisé* [8]

$$\beta_{n+1} \in \operatorname{argmin}_\beta c(\alpha_n, \beta) - F^c(\beta) \quad (18a)$$

$$\alpha_{n+1} \in \operatorname{argmin}_\alpha c(\alpha, \beta_{n+1}). \quad (18b)$$

théorème 5. *Soit $c \colon X \times Y \to \mathbf{R}$ un coût continu entre deux compacts X et Y, et F une fonction c-concave. L'algorithme* (18) *s'écrit*

$$\beta_{n+1} \in \partial_c F(\alpha_n) \quad (19a)$$

$$\alpha_{n+1} \in \operatorname{argmin}_\alpha c(\alpha, \beta_{n+1}), \quad (19b)$$

ce qui garantit une propriété de descente

$$F(\alpha_{n+1}) - F(\alpha_n) \leq c(\alpha_{n+1}, \beta_{n+1}) - c(\alpha_n, \beta_{n+1}) \leq 0. \quad (20)$$

Démonstration. Par application des théorèmes 1 pour la forme des itérés, et 2 pour leur existence. Ce théorème généralise [8, 3.3, 3.15(i)] aux fonctions non différentiables, avec l'existence par le théorème de Weierstrass. Prouvons la propriété de descente : puisque $\beta_{n+1} \in \partial_c F(\alpha_n)$ on peut appliquer l'équation (7) à α_{n+1} pour obtenir $F(\alpha_{n+1}) - F(\alpha_n) \leq c(\alpha_{n+1}, \beta_{n+1}) - c(\alpha_n, \beta_{n+1})$. L'étape (18b) assure que cette dernière quantité est négative. □

4 Descente de Gradient par transport

On vient de mettre en place un cadre algorithmique pour la minimisation des fonctions c-concaves. Dans cette section, on applique ce cadre au coût de transport $\mathcal{T}_c \colon P(X) \times P(X')$; l'algorithme utilise alors le coût $\mathcal{T}_c$ qui dépend à son tour de la fonction de couplage c.

théorème 6. *Soit c un coût continu entre deux compacts X et Y, $a \in P(X)$ et $F \colon P(X) \to \mathbf{R}$ une fonction convexe.*

$$\partial_{\mathcal{T}_c} F(\alpha) \neq \emptyset \implies \partial F(\alpha) \subset \Gamma_c(X). \quad (21)$$

où $f \in \Gamma_c(X)$ est une fonction c-concave sur le support de α.

Démonstration. Par le théorème 1, $\beta \in \partial_{\mathcal{T}_c} F(\alpha)$ si et seulement si $\alpha \in \operatorname{argmin}_{\alpha'}\{\mathcal{T}_c(\alpha', \beta) - F(\alpha')\}$. Puisque F et $G_\beta = \alpha \mapsto \mathcal{T}_c(\alpha, \beta)$ sont convexes, c'est aussi équivalent à l'inclusion $\partial^\epsilon F(\alpha) \subset \partial^\epsilon G_\beta(\alpha)$ des ϵ-sous-différentiels pour tout $\epsilon \geq 0$ [5]. Les ϵ-sous-gradients de G_β sont les potentiels optimaux à ϵ près pour le dual (3) ([11, 7.17]), donc en particulier $\partial F(\alpha) \subset \partial G_\beta(\alpha)$ implique que les sous-gradients de F à α coïncident avec une fonction c-concave sur le support de α. □

La réciproquement s'obtient au moins dans un cas simple.

proposition 7. *Soit $F\colon P(X) \to \mathbf{R},\ \alpha \mapsto \langle \alpha, f\rangle + c$ une fonction linéaire. $\partial_{\mathcal{T}_c} F(\alpha) \neq \emptyset \iff f \subset \Gamma_c(X)$.*

Démonstration. La fonction $\alpha \mapsto \mathcal{T}_c(\alpha, \beta) - F(\alpha)$ est maintenant convexe, donc les conditions d'optimalité au premier ordre sont suffisantes. Ce cadre est étudié dans [10]. □

La description des sous-différentiels rend l'algorithme (19) explicites.

théorème 8. *Soit $\mathcal{F}(h)$ la fonction de couplage générée par une fonction de type Legendre $h\colon X \to \mathbf{R}$. On suppose que $F\colon P(X) \to \mathbf{R}$ est une fonction convexe et différentiable, avec un gradient $\nabla F(\alpha)\colon X \to \mathbf{R}$ h-lisse et différentiable pour tout α, et que $\partial_{\mathcal{T}_c} F(\alpha) \neq \emptyset$. La descente de gradient généralisée avec le coût $\mathcal{T}_c$ s'écrit*

$$\beta_{n+1} = \nabla(h - \nabla F(\alpha_n))\#\alpha_n \tag{22a}$$
$$\alpha_{n+1} = \nabla(h^*)\#\beta_{n+1}. \tag{22b}$$

Démonstration. L'implication du théorème 6 montre que $\nabla F(\alpha)$ est un potentiel optimal dans le transport de α_n vers β_{n+1}. Par le théorème 4 et par différentiabilité, il existe un plan optimal π concentré sur $\partial(h - \nabla F(\alpha_n)) = \{\nabla(h - \nabla F(\alpha_n))\}$. On vérifie ensuite que $\mathcal{T}_c(\alpha, \nabla h\#\alpha) = 0$, ce qui est son minimum, et on obtient (22b) puisque ∇h est une bijection d'inverse ∇h^*. □

Le théorème 8 montre que, si les itérés existent, alors choisir α_{n+1} comme mesure image de α_n par le pas de miroir

$$\nabla(h^*) \circ \nabla(h - \nabla F(\alpha_n)) \tag{23}$$

réalise les itérations (18) pour le coût $\mathcal{T}_c$.

théorème 9. *Sous les mêmes hypothèses que le théorème 8, pour $T_n = \nabla(h - \nabla F(\alpha_n))$, le progrès minimal* (20) *s'écrit*

$$F(\alpha_{n+1}) - F(\alpha_n) \leq -\int \mathcal{F}(h)(x, T_n(x))\,\mathrm{d}\alpha_n(x). \tag{24}$$

Démonstration. Puisque β_{n+1} est choisi tel que l'application $T_n = \nabla(h - \nabla F(\alpha_n))$ transporte la mesure α_n sur β_{n+1}, on peut calculer le progrès de l'étape n par la formule de Monge $\mathcal{T}_c(\alpha_n, \beta_{n+1}) = \int \mathcal{F}(h)(x, T_n(x))\,\mathrm{d}\alpha_n(x)$. □

Pour le coût quadratique ($h = \frac{1}{2}\|\cdot\|^2$) la décroissance suffisante s'écrit

$$F(\alpha_{n+1}) - F(\alpha_n) \leq -\frac{1}{2}\int \|\nabla F(\alpha_n)\|^2\,\mathrm{d}\alpha_n. \tag{25}$$

On retrouve bien le progrès minimal pour la discrétisation du flot de gradient en Wasserstein [3, 2.5, $\alpha = 0$].

Recherche Linéaire. Prenons une taille de pas $\lambda > 0$ et son coût associé $\lambda\mathcal{T}_c$; cela revient à faire les itérations données par $\nabla(h^*) \circ \nabla(h - \lambda^{-1}\nabla F(\alpha_n))$. L'hypothèse de concavité du coût $\mathcal{T}_c$ étant trop restrictive, on la remplace par une hypothèse de décroissance locale, compatible avec la propriété de descente (20). Concrètement, cela signifie qu'à chaque itération α_n, une recherche linéaire est effectuée pour déterminer un pas admissible λ_n (dépendant maintenant de l'itération) garantissant la descente (20).

5 Application au BLASSO

On applique l'algorithme (22) à la décomposition d'un signal $y \in \mathbf{R}^m$ comme la somme

$$y = \sum_{i=1}^{n} a_i \phi(x_i) \tag{26}$$

d'éléments $\phi(x_1), \ldots, \phi(x_n)$ avec des poids $a_1, \ldots, a_n$ positifs. On suppose que le dictionnaire $\phi\colon X \to \mathbf{R}^m$ est une fonction différentiable avec une dérivée L-Lipschitz. On cherche à l'estimer comme le minimiseur de $F + G$

$$\min_{\alpha \in M(X)} F(\alpha) + G(\alpha) = (1/2)\,\|\int_X \phi\,\mathrm{d}\alpha - y\|^2 + \rho|\alpha|(X) \tag{27}$$

où $G(\alpha) = |\alpha|(X)$ est la variation totale de la mesure α. Ce problème, appelé BLASSO (Beurling-LASSO), étends le cadre du LASSO à l'espace des mesures — on verra [7] pour une synthèse récente.

Pour minimiser le BLASSO, on implémente le pas miroir (23) sur les positions et un pas de gradient proximal par seuillage doux sur les poids. En pratique, la mesure α_n sera approximée numériquement par un système de particules. Il est nécessaire de vérifier a minima les hypothèses du théorème 6 afin d'assurer la validité de l'approche.

proposition 10. *Quand ϕ est lisse, La fonction F vérifie les conditions nécessaires 6 pour le coût quadratique.*

Démonstration. En tant que fonction sur $M(X)$, F a pour gradient $\nabla F(\alpha)\colon X \to \mathbf{R},\ x \mapsto \langle \phi(x), \int_X \phi\,\mathrm{d}\alpha - y\rangle_{\mathbf{R}^m}$. Pour tout vecteur de résidus $r \in \mathbf{R}^m$, la fonction $x \mapsto D\phi(x) \cdot r$ est Lipschitz, donc le gradient $\nabla F(\alpha)$ est aussi Lipschitz. □

Application à la localisation de sources. Chaque source émet un signal donné par $\phi(x)_i = \mathrm{sinc}_2((t_i - x)/10.0)$ où $\mathrm{sinc}_2\colon \mathbf{R}^2 \to \mathbf{R}$ est défini comme le produit sur chaque axe du sinus cardinal $\mathrm{sinc}(x) = \frac{\sin \pi x}{\pi x}$ et $(t_i)_{i=1}^m \in [0,1]^2$ sont les positions des capteurs. On observe un signal non bruité $y \in \mathbf{R}^m$ qui est produit par la superposition de 8 sources aléatoirement choisies sur $[0,1]^2$. Les $m = 12 \times 12$ capteurs sont disposés selon une grille régulière sur $[0,1]^2$. Le choix de l'entropie négative

$$h : x \mapsto \sum_{i=1}^{2} x_i \log x_i + (1 - x_i)\log(1 - x_i) \tag{28}$$

avec comme miroir associé $\nabla h^*(x)_i = \frac{1}{1+\exp(-x_i)}$ permet de rester dans le domaine $[0,1]^2$ où se trouvent les sources. Le choix de l'exemple est assez favorable à l'implémentation

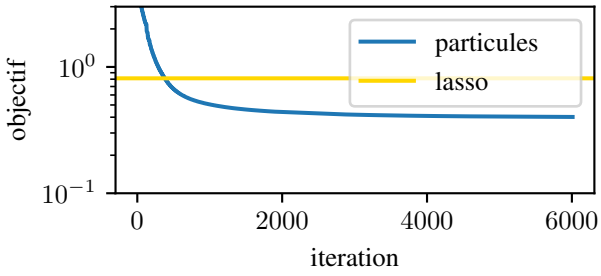

FIGURE 1 : Comparaison entre la valeur de l'objectif (27) pour le LASSO (après convergence) et pour les itérations de l'algorithme proposé.

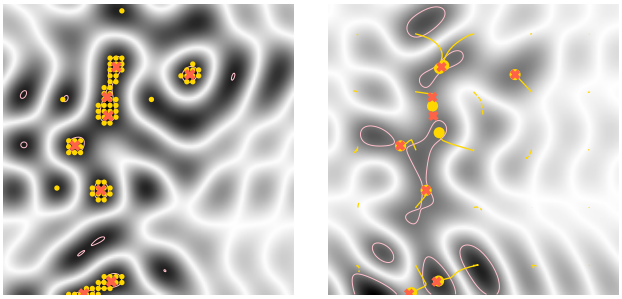

FIGURE 2 : Le support obtenu par le LASSO (à gauche) contre celui obtenu par particules. En niveau de gris est représenté la valeur absolue du gradient $|\nabla F|$, en jaune les trajectoires des particules et leur limites. En croix orange, le support de la mesure source.

numérique de la méthode proposée. En effet, le lobe principal du sinus cardinal est suffisamment large pour attirer les trajectoires des particules.

On compare, avec un taux de régularisation à $\rho = 0.05$, l'algorithme du LASSO sur une grille fixe de 50×50 et l'algorithme proposé initialisé par une mesure discrète supportée sur une grille régulière de 5×5.

On peut observer sur la figure 2 les trajectoires des particules, et en pointillés les trajectoires avec un poids nul à la dernière itération. Le gradient est tracé sur une grille fine, ainsi que la ligne de niveau $|\nabla F(x)| = \rho$.

On vérifie sur la figure 1 que l'algorithme proposé est bien un algorithme de descente, et qu'il obtient une meilleure valeur de l'objectif que si l'on fixe les positions sur une grille fine. Cet exemple soutient l'idée selon laquelle une méthode adaptative, capable de prendre en compte la régularité de l'objectif — ici sous la forme de la c-concavité — constitue une alternative efficace à l'approche par grille, notamment en dimension élevée.

6 Conclusion

Il sera pertinent de discuter de la discrétisation par système de particules, ainsi la prise en compte des poids en plus des positions. Pour l'algorithme de descente sur les mesures, on pourra s'intéresser à un coût quelconque plutôt que miroir : la structure de l'étude sera sensiblement similaire.

Finalement, nous avons présenté un algorithme pour la minimisation de fonctions convexes sur les mesures de probabilité, ainsi que des conditions nécessaires à sa mise œuvre dans le cadre de la descente de gradient généralisée. Une application numérique illustre la pertinence de la recherche linéaire dans les cas pratique. Cet algorithme est conçu comme une méthode itérative explicite, contrairement à d'autres approches comme le JKO [6] qui est un schéma d'Euler implicite.

Références

[1] Mathieu BLONDEL, André F. T. MARTINS et Vlad NICULAE : Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35), 2020.

[2] Nicholas BOYD, Geoffrey SCHIEBINGER et Benjamin RECHT : The Alternating Descent Conditional Gradient Method for Sparse Inverse Problems. *SIAM Journal on Optimization*, 27(2), 2017.

[3] Lénaïc CHIZAT : Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 194(1), 2022.

[4] Quentin DENOYELLE, Vincent DUVAL, Gabriel PEYRÉ et Emmanuel SOUBIES : The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1), 2020.

[5] J.-B. HIRIART-URRUTY : From Convex Optimization to Nonconvex Optimization. Necessary and Sufficient Conditions for Global Optimality. *In Nonsmooth Optimization and Related Topics*. Springer US, 1989.

[6] Richard JORDAN, David KINDERLEHRER et Felix OTTO : The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1), 1998.

[7] Bastien LAVILLE, Laure BLANC-FÉRAUD et Gilles AUBERT : Off-The-Grid Variational Sparse Spike Recovery : Methods and Algorithms. *Journal of Imaging*, 7(12), 2021.

[8] Flavien LÉGER et Pierre-Cyril AUBIN-FRANKOWSKI : Gradient descent with a general cost, 2023.

[9] Jean Jacques MOREAU : Inf-convolution, sous-additivité, convexité des fonctions numériques. *Journal de Mathématiques Pures et Appliquées*, 1970.

[10] Adil SALIM, Anna KORBA et Giulia LUISE : The Wasserstein Proximal Gradient Algorithm. *In Advances in Neural Information Processing Systems*, volume 33, 2020.

[11] Filippo SANTAMBROGIO : *Optimal Transport for Applied Mathematicians : Calculus of Variations, PDEs, and Modeling*. Springer International Publishing, 2015.

[12] Walter SCHACHERMAYER et Josef TEICHMANN : Characterization of optimal transport plans for the Monge-Kantorovich problem. *Proceedings of the American Mathematical Society*, 137(2), 2009.