Accélération parcimonieuse des poids exponentiels

Pierre Gaillard [1] & Olivier Wintenberger [2]

[1] pierre.gaillard@inria.fr
INRIA - Sierra project-team, Département d'Informatique de l'Ecole Normale Supérieure, Paris, France
[2] olivier.wintenberger@upmc.fr
Sorbonne Universités, UPMC Univ Paris 06 LSTA, FRANCE

Résumé. On considère le problème d'optimisation séquentielle d'une suite indépendente et identiquement distribuée (i.i.d.) de n fonctions convexes en observant leurs gradients. Dans ce travail, on introduit une procédure permettant d'accélérer la convergence de procédures lentes en $\mathcal{O}(1/\sqrt{n})$ en vitesse rapide $\mathcal{O}(1/n)$ quand le risque est fortement convexe. Si le minimiseur du risque $\theta^* \in \mathbb{R}^d$ est parcimonieux $\|\theta^*\|_0 = d_0 \ll d$, cela permet en particulier de dépendre linéairement de d_0 plutôt que de d. Cette communication est basée sur un article récemment publiée par les mêmes auteurs [1].

Mots-clés. Apprentissage séquentiel, Parcimonie, Optimisation

Abstract. We consider the sequential optimization of an i.i.d. sequence of n convex functions by observing their gradients. This work intends to accelerate the slow convergence rate of some procedures in $\mathcal{O}(1/\sqrt{n})$ into the fast rate $\mathcal{O}(1/n)$ when the risk is strongly convex. If the risk minimizer $\theta^* \in \mathbb{R}^d$ is sparse (i.e., $\|\theta^*\|_0 = d_0 \ll d$) the rate depends linearly on d_0 instead of d. This communication is based on a recently published paper by the same authors [1].

Keywords. Online learning, sparsity

1 Introduction

On considère le problème où une suite i.i.d. de fonctions convexes $\ell_1, \dots, \ell_n : \mathbb{R}^d \to R$ sont optimisées séquentiellement en observant leurs gradients $\nabla \ell_t$. L'objectif est à chaque instant $t \geq 1$, de construire un estimateur $\bar{\theta}_{t-1}$ minimisant le risque $\theta \in \mathbb{R}^d \mapsto \mathbb{E}[\ell_t](\theta)$ à partir de l'information passée seulement (i.e., les gradients observés jusqu'à l'instant $t-1$). Sans hypothèse supplémentaire, des procédures de descente de gradient [2, 5] garantissent une vitesse optimale $\mathcal{O}\big(\sqrt{\log(d)/t}\big)$ [1]. Si le risque est fortement convexe, des descentes de gradient atteignent une vitesse plus rapide de l'ordre $\mathcal{O}(d/t)$ également optimale. On présente ici une procédure d'accélération SAEW qui accélère les premiers algorithmes

1. La notation $\mathcal{O}$ cache des constantes indépendantes de t et d

dans le cas fortement convexe. Si le paramètre optimal θ^* est de plus parcimonieux de dimension $d_0 := \|\theta^*\|_0 \ll d$, nous obtenons la vitesse $\mathcal{O}\big(d_0 \log(d)/t\big)$. Une telle vitesse est similaire à celle que peut obtenir le Lasso dans un cadre non séquentiel avec design aléatoire. Remarquons que celle-ci n'est atteignable en temps polynomial qu'au prix d'un facteur multiplicatif α^{-1}, où α est le paramètre de forte convexité (voir [4]).

Hypothèses Tout au long du papier, nous faisons les hypothèse suivantes. Nous supposons que le risque admet un unique minimiseur $\theta^* \in \mathbb{R}^d$ et qu'il existe $r \geq 2$ et $\alpha > 0$ tels que

$$\forall \theta \in \mathbb{R}^d \qquad \alpha \big\|\theta - \theta^*\big\|_2^r \leq \mathbb{E}\big[\ell_t\big](\theta) - \mathbb{E}\big[\ell_t\big](\theta^*)\,. \tag{1}$$

Remarquons que cette hypothèse est impliquée par la forte convexité avec $r = 2$. Nous supposons de plus que $\|\theta^*\|_1 \leq U$, $\|\theta^*\|_0 \leq d_0$ et que les gradients sont bornés localement en norme infinie $\max_{\theta:\|\theta\|_1 \leq 2U} \|\nabla \ell_t(\theta)\|_\infty \leq B$.

2 Algorithme et résultat

SAEW est basé sur un algorithme de base, par exemple l'algorithme BOA [3], qu'il réinitialise régulièrement pour optimiser $\mathbb{E}[\ell_t]$ dans des boules ℓ_1 exponentiellement plus petites. À chaque session $i \geq 0$, l'algorithme BOA est appliqué dans $\mathcal{B}_1(\hat{\theta}_{i-1}, U_i)$ la boule ℓ_1 centrée en l'estimateur parcimonieux $\hat{\theta}_{i-1}$ et de rayon $U_i = U2^{-i}$. Après t_i pas de temps, ce dernier forme un estimateur $\bar{\theta}_i$ dont le risque satisfait (voir [3]) avec grande probabilité

$$\mathbb{E}[\ell_t](\bar{\theta}_i) - \mathbb{E}[\ell_t](\theta^*) \lesssim U_i \sqrt{\frac{\log d}{t_i}}\,, \tag{2}$$

où $\lesssim$ cache des constantes pouvant dépendre de B, de la probabilité δ que la borne ne soit pas satisfaite ou de termes en $\log\log$. L'idée est de choisir t_i le temps d'arrêt de la i-ème session pour que l'on puisse construire à partir de l'inégalité (2) l'estimateur $\hat{\theta}_i$ tel que $\|\hat{\theta}_i - \theta^*\|_1 \leq U_i/2 =: U_{i+1}$ avec grande probabilité. On réinitialise alors l'algorithme pour qu'il optimise dans la boule ℓ_1 centrée en $\hat{\theta}_i$ et de rayon U_{i+1}. La procédure est illustrée en Figure 1. Plus précisément, l'algorithme définit

$$t_i \approx 2^{2i(r-1)} \frac{d_0^r \log(d)}{\alpha^2 U^{2(r-1)}} \qquad \text{et} \qquad \hat{\theta}_i := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d : \|\theta\|_0 \leq d_0} \|\theta - \bar{\theta}_i\|_2\,, \tag{3}$$

où le signe $\approx$ cache des constantes.

Théorem 1. *Soit $n \geq 1$ et soit i_n tel que $\sum_{i=1}^{i_n} t_i \leq n \leq \sum_{i=1}^{i_n+1} t_i$. Alors, l'algorithme décrit ci-dessus qui effectue BOA réinitialisé tous les t_i pas de temps sur la boule $\mathcal{B}_1(\hat{\theta}_{i-1}, U2^{-i})$ (où t_i et $\hat{\theta}_i$ sont définis en (3)) satisfait*

$$\mathbb{E}[\ell_t](\bar{\theta}_{i_n}) - \mathbb{E}[\ell_t](\theta^*) \lesssim \alpha^{-\frac{1}{r-1}} \left(\frac{d_0 \log d}{n}\right)^{\frac{r}{2(r-1)}}\,.$$

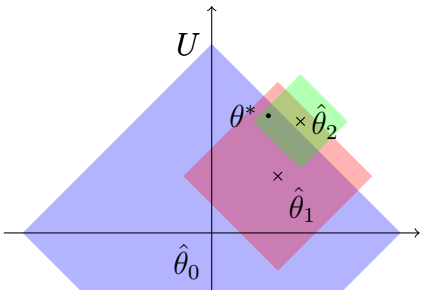

FIGURE 1 – La procédure d'accélération. L'algorithme commence par optimiser dans la boule bleue, puis dans la rouge, la verte,...

On présente ici une esquisse de preuve. La preuve rigoureuse est disponible dans [1] dans le cas particulier de la forte convexité (i.e., $r = 2$). Dans ce cas, la vitesse est de l'ordre $\mathcal{O}\big(d_0 \log(d)/n\big)$.

Esquisse de preuve. Commençons par montrer que les choix $\hat{\theta}_i$ et t_i définis en (3) satisfont

$$\|\hat{\theta}_i - \theta^*\| \leq U 2^{-(i+1)} \,. \tag{4}$$

Soit $i \geq 0$. Après t_i pas de temps, BOA satisfait d'après l'inégalité (2),

$$\mathbb{E}[\ell_t](\bar{\theta}_i) - \mathbb{E}[\ell_t](\theta^*) \lesssim U_i \sqrt{\frac{\log d}{t_i}} \,. \tag{5}$$

En utilisant l'hypothèse (1), cela donne

$$\|\bar{\theta}_i - \theta^*\|_2 \lesssim \left(\frac{U_i}{\alpha} \sqrt{\frac{\log d}{t_i}} \right)^{1/r} \,. \tag{6}$$

Comme θ^* a aussi seulement d_0 coordonnées non nulles, par définition (3) de $\hat{\theta}_i$, on a $\|\hat{\theta}_i - \bar{\theta}_i\|_2 \leq \|\bar{\theta}_i - \theta^*\|_2$. De plus,

$$\big\|\hat{\theta}_i - \theta^*\big\|_1 \leq \sqrt{2d_0} \big\|\hat{\theta}_i - \theta^*\big\|_2 \leq \sqrt{2d_0} \Big(\big\|\hat{\theta}_i - \bar{\theta}_i\big\|_2 + \big\|\bar{\theta}_i - \theta^*\big\|_2 \Big) \leq 2\sqrt{2d_0} \big\|\bar{\theta}_i - \theta^*\big\|_2 \,.$$

D'après (6), on obtient

$$\big\|\hat{\theta}_i - \theta^*\big\|_1 \lesssim \sqrt{d_0} \left(\frac{U_i}{\alpha} \sqrt{\frac{\log d}{t_i}} \right)^{1/r} \,.$$

On choisit alors t_i tel que

$$\sqrt{d_0}\left(\frac{U_i}{\alpha}\sqrt{\frac{\log d}{t_i}}\right)^{1/r} \approx \frac{U_i}{2} \Leftrightarrow t_i \approx 2^{2i(r-1)}\frac{d_0^r \log(d)}{\alpha^2 U^{2(r-1)}}\,,$$

où on a utilisé que $U_i = U2^{-i}$. Ceci conclut la preuve de (4). On peut alors commencer la nouvelle session d'optimisation dans la boule centrée en $\hat{\theta}_i$ et de rayon $U_{i+1} = U2^{-(i+1)}$.

Soit i_n le nombre de sessions terminées avant n pas de temps. L'estimateur de BOA lors de la dernière session $\bar{\theta}_{i_n}$ satisfait l'équation (5). Il reste à évaluer U_{i_n}. Les t_i augmentant exponentiellement, on en déduit que la somme des temps t_i de chaque session jusqu'à la session i_n est approximativement celle de la dernière $t_{i_n} \lesssim n \lesssim t_{i_n}$. On en déduit

$$n \lesssim 2^{2i_n(r-1)}\frac{d_0^r \log(d)}{\alpha^2 U^{2(r-1)}} \Rightarrow U_{i_n} := U2^{-i_n} \lesssim \alpha^{-\frac{1}{r-1}}\left(\frac{d_0^r \log d}{n}\right)^{\frac{1}{2(r-1)}}\,.$$

Réinjecter dans (5) conclut la preuve

$$\mathbb{E}[\ell_t](\bar{\theta}_{i_n}) - \mathbb{E}[\ell_t](\theta^*) \lesssim \alpha^{-\frac{1}{r-1}}\left(\frac{d_0 \log d}{n}\right)^{\frac{r}{2(r-1)}}\,.$$

□

Dans l'exposé, on présentera également des applications à la régression linéaire, la régression quantile, ainsi que des simulations. De plus, pour calibrer les temps de réinitialisation t_i, l'algorithme a besoin de connaître à l'avance les valeurs de U, d_0, α, r et B. On précisera dans la présentation comment calibrer ces paramètres séquentiellement en utilisant une surcouche d'agrégation.

Références

[1] P. Gaillard et O. Wintenberger. "Sparse Accelerated Exponential Weights". Accepted at AISTAT'17. 2017.

[2] J. Kivinen et M. K. Warmuth. "Exponentiated Gradient Versus Gradient Descent for Linear Predictors". In : *Information and Computation* 132.1 (1997), p. 1–63.

[3] O. Wintenberger. "Optimal learning with Bernstein Online Aggregation". In : Extended version available at arXiv :1404.1356 [stat. ML] (2014).

[4] Y. Zhang, M. J. Wainwright et M. I. Jordan. "Lower bounds on the performance of polynomial-time algorithms for sparse linear regression." In : *COLT*. 2014, p. 921–948.

[5] M. Zinkevich. "Online Convex Programming and Generalized Infinitesimal Gradient Ascent". In : *Proceedings of the 20th International Conference on Machine Learning, ICML 2003*. 2003.