Descente de gradient stochastique moyen

Jérémy Bettinger

Article de Francis Bach, ENS Paris, Journal of Machine Learning (2014)

Table des matières

1 Résumé **1**

2 Introduction **2**

3 Hypothèses **3**

4 Résultats antérieurs **4**

5 Inégalités de concentration **4**

6 Convergence des gradients **5**

7 Utilisation avancée de l'autoconcordance **6**

1 Résumé

Méthode de descente de gradient classique : Calculer le gradient, adapter, répéter la procédure...

Problème : Quand on a beaucoup de données le calcul exact de gradient (de la fonction de perte) est long et lourd.

Solution : Gradient Stochastique : C'est une estimation du vrai gradient via estimation sur des petites données. 1) on tire des données au hasard 2) on utilise ces données avec un gradient 3) on itère la procédure.

Ce que l'on va faire : Étudier des problèmes d'apprentissage supervisé comme la régression logistique dans le cadre où les observations ne sont utilisées qu'une fois.

On va montrer qu'après N itérations, avec une taille de pas proportionelle à $1/R^2\sqrt{N}$ la vitesse de convergence est de l'ordre de $O(1/\sqrt{N})$ où : N désigne le nombre d'observations $R =$ le maximum des normes des observations et peut aller jusqu'à $O(R^2/(\mu N))$ où μ désigne la plus petite valeur propre de la Hessienne au point extremal.

On montrera que puisque μ n'a pas besoin d'être connu à l'avance, cela montre que le gradient stochastique moyen s'adapte à la convexité forte, inconnue, locale de la fonction à estimer.

La démonstration reposera sur des propriétés d'autoconcordance générales de la perte logistique.

2 Introduction

La minimisation d'une fonction qui n'est accessible que par des estimations non biaisées de la fonction ou de ses gradients est un problème clé dans beaucoup de domaines des maths.

Pour des problèmes d'optimisation convexe, la vitesse de convergence dépend beaucoup de la (forte) convexité de la fonction à étudier.

On rappelle la définition de fonction fortement convexe :

Définition 2.1. *f est fortement convexe de module $\mu > 0$ si on a*

$$\forall x, y \in E \quad f(tx + (1-t)y) \leqslant tf(x) + (1-t)f(y) - \mu t(1-t)\|x-y\|^2/2.$$

Pour des fonctions μ fortement convexes après n itérations le taux optimal de convergence est en $O(1/\mu n)$ et pour les fonctions convexes en $O(1/\sqrt{n})$ via une méthode de gradient stochastique avec pas proportionnels à $1/\mu n$ et $1/\sqrt{n}$ respectivement.
Pour les fonctions régulières avec pas proportionnel à $1/\sqrt{n}$ se comporte en $O((\log n)^\gamma/\sqrt{n})$.

Les problèmes d'optimisation convexes sont de la forme $f(\theta) = \mathbb{E}(\ell(t, \langle\theta, x\rangle)$ où ℓ est la fonction de perte, $x \in \mathcal{H}$ un espace de Hilbert et où $\theta \in \mathcal{H}$ est la prédiction linéaire.

On a ou pas la convexité selon les cas, souvent cela dépend de :
a) de la corrélation entre les données (x_i).
b) de la (forte) convexité de la fonction de coût ℓ.

La fonction de perte logistique $\ell(u) = \ln(1 + e^{-u})$ n'est pas fortement convexe, à moins de se restreindre à un compact K car $\ell''(u) = e^{-u}(1 + e^{-u})^{-2} \geqslant c_K$.

On ne peut donc pas appliquer les résultats sur la forte convexité.

Objectifs : Montrer qu'avec des hypothèses appropriées, à savoir l'autoconcordance on peut avoir une vitesse de convergence en $O(R^2/\mu n)$ où μ est la plus petite valeur propre de la Hessienne à l'extremum global sans prendre les termes en $e^{\alpha u}, \alpha > 0$.

Un autre but est de trouver une méthode qui permette de se passer de la connaissance de la constante de forte convexité locale.
Dans des situations régulières, le gradient stochastique moyenné à pas $O(1/\sqrt{n})$ est adapté à la forte convexité du problème.

Mais cela ne s'applique pas ici car même dans les contextes de faibles dépendances, on n'a pas toujours, dans le cadre de la régression logistique, la forte convexité globale.

A suivre : Gradient stochastique moyen pour la régression logistique avec un pas proportionnel à $1/R^2\sqrt{n}$. On va montrer que l'algorithme s'adapte à la constante locale de la forte convexité i.e. la plus petite valeur propre de la hessienne à l'extremum.

On se fixe N et un pas constant décroissant en N comme $1/R^2\sqrt{N}$.

Remarque : On pourrait changer ce pas en cours de route avec une astuce de "doublement".

3 Hypothèses

Soit f une fonction définie sur un espace de Hilbert $(\mathcal{H}, \|.\|)$ on identifie l'espace $\mathcal{H}$ avec son dual. Ainsi les gradients de f sont aussi dans $(\mathcal{H}, \|.\|)$.

On se prend une filtration $(\mathcal{F}_n)_n, \theta_0 \in \mathcal{H}, f_n : \mathcal{H} \to \mathbb{R}$, pour $R > 0$ on suppose les hypothèses suivantes :

(A1) (régularité) : f est convexe et 3 fois différentiable.

(A2) (autoconcordance) : $\forall \theta_1, \theta_2 \in \mathcal{H},\ \phi : t \mapsto f(\theta_1 + t(\theta_2 - \theta_1))$ vérifie $|\phi'''(t)| \leqslant R\|\theta_1 - \theta_2\|\phi''(t)$.

(A3) (minimum global) : f a un minimum global en $\theta_\star \in \mathcal{H}$.

(A4) (bornitude lipschitz) : $\|f'(\theta)\| \leqslant R \ \ ; \ \ \|f_n'(\theta)\| \leqslant R$.

(A5) (adaptabilité) : f_n est $\mathcal{F}_n-$mesurable.

(A6) (gradient non biaisé) : $\mathbb{E}(f_n'(\theta_{n-1})|\mathcal{F}_{n-1}) = f'(\theta_{n-1})$.

(A7) (récurrence du gradient stochastique) : $\theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_{n-1})$ où $(\gamma_n)_n$ est une suite déterministe.

On pose $\overline{\theta_n} = \frac{1}{n}\sum_{k=0}^{n} \theta_k$ ce qui donne $\overline{\theta_n} = \theta_{n-1}/n + \overline{\theta_{n-1}}(n-1)/n$.

La condition **(A2)** d'autoconcordance est standard en optimisation mais ce n'est pas exactement la même. Bach a généralisé cela.

Exemples de fonctions vérifiant ces hypothèses :

*Régression logistique : $f_n(\theta) = \log(1 + \exp(-y_n\langle x_n, \theta\rangle))$, (x_n) p.s. uniformément bornés par R et $y_n \in \{-1, 1\}$.

*Modèle linéaire généralisé : $f_n(\theta) = -\langle\theta, \psi(x_n, y_n)\rangle + \log\int h(y)\exp(\langle\theta, \psi(x_n, y)\rangle)\,\mathrm{d}y$ avec la quantité $\psi(x_n, y) \in \mathcal{H}$ p.s. borné en norme par R.

Complexité en $O(dn)$ si $dim(H) = d$. Plus compliqué si $dim(H) = +\infty$.

4 Résultats antérieurs

Proposition 4.1. *On suppose ici f lipschitz non fortement convexe. Si on prend $\gamma_n = \gamma > 0$ et qu'on suppose* ***(A1)-(A7) \ (A2)*** *alors :*

$$\mathbb{E}(f(\overline{\theta_n})) - f(\theta_\star) + \frac{1}{2\gamma n}\mathbb{E}\|\theta_n - \theta_\star\|^2 \leqslant \frac{1}{2\gamma n}\|\theta_0 - \theta_\star\|^2 + \frac{\gamma}{2}R^2$$

donc

$$\mathbb{E}(f(\overline{\theta_n}) - f(\theta_\star)) = O\left(\gamma + \frac{1}{\gamma n}\right)$$

on égalise les vitesses, et on prend γ proportionnel à $1/\sqrt{n}$.

Si on prend $\gamma = 1/(2R^2\sqrt{N})$:

$$\mathbb{E}\|\theta_n - \theta_\star\|^2 \leqslant \|\theta_0 - \theta_\star\|^2 + \frac{1}{4R^2}$$

et

$$\mathbb{E}(f(\overline{\theta_N})) - f(\theta_\star) \leqslant \frac{R^2}{\sqrt{N}}\|\theta_0 - \theta_\star\|^2 + \frac{1}{4\sqrt{N}}.$$

Si on connaît $\|\theta_0 - \theta_\star\|$ on pose $\gamma = \|\theta_0 - \theta_\star\|/(R\sqrt{N})$ pour avoir une vitesse de convergence proportionnelle à $R\|\theta_0 - \theta_\star\|/\sqrt{N}$.

Problème : On ne connaît pas toujours bien $\|\theta_0 - \theta_\star\|$ ni une majoration facile...

Remarque : La borne ne dépend pas explicitement de la dimension, mais en pratique la quantité $R^2\|\theta_0 - \theta_\star\|^2$ croit avec la dimension.

But de la suite :

1) On va montrer que les moments de $\|\theta_n - \theta_\star\|^2$ et de $f(\overline{\theta_N}) - f(\theta_\star)$ sont bornés. Cela va nous fournir un comportement sous exponentiel.

2) On va montrer que la norme au carré du gradient en $\overline{\theta_N}$ converge en $O(1/n)$ dans le cas non fortement convexe. Cela nous permettra de trouver une inégalité de concentration sur le gradient, ce qui nous permettra de pourvoir appliquer un lemme technique pour les fonctions autoconcordantes et fortement convexes localement.

5 Inégalités de concentration

Proposition 5.1. *Sous les hypothèses* ***(A1)-(A7) \ (A2)****, avec un pas égal à γ on a pour $p \geqslant 1$:*

$$\mathbb{E}\left(2\gamma n[f(\overline{\theta_n}) - f(\theta_\star)] + \|\theta_n - \theta_\star\|^2\right)^p \leqslant (3\|\theta_0 - \theta_\star\|^2 + 20np\gamma^2R^2)^p.$$

Corollaire 5.2. *Avec $\gamma = 1/(2R^2\sqrt{N})$ on a :*

$$\mathbb{E}\|\theta_n - \theta_\star\|^{2p} \leqslant \left(\frac{1}{R^2}(3R^2\|\theta_0 - \theta_\star\|^2 + 5p)\right)^p,$$

$$\mathbb{E}\|f(\overline{\theta_n} - f(\theta_\star)\|^p \leqslant \left(\frac{1}{\sqrt{N}}(3R^2\|\theta_0 - \theta_\star\|^2 + 5p)\right)^p.$$

Démonstration. Soit utiliser une relation de récurrence et utiliser une inégalité de moments martingale. Soit utiliser une inégalité type Burkholder. □

Maintenant qu'on a les moments (dont les normes p sont plus petites qu'une fonction affine en p d'où un comportement sous exponentiel), on obtient via une inégalité de concentration :

Proposition 5.3. *Sous les hypothèses **(A1)-(A7) \ (A2)**, avec un pas égal à γ on a pour $t \geqslant 0$:*

$$\mathbb{P}\left(f(\overline{\theta_n}) - f(\theta_\star) \geqslant 30\gamma R^2 t + \frac{3\|\theta_0 - \theta_\star\|^2}{\gamma n}\right) \leqslant 2\exp(-t),$$

$$\mathbb{P}\left(\|\theta_n - \theta_\star\|^2 \geqslant 60n\gamma^2 R^2 t + 6\|\theta_0 - \theta_\star\|^2\right) \leqslant 2\exp(-t).$$

Corollaire 5.4. *Sous les hypothèses **(A1)-(A7) \ (A2)**, avec un pas égal à $\gamma = 1/(2R^2\sqrt{N})$ on a pour $t \geqslant 0$:*

$$\mathbb{P}\left(f(\overline{\theta_n}) - f(\theta_\star) \geqslant \frac{15t}{\sqrt{N}} + \frac{5R^2\|\theta_0 - \theta_\star\|^2}{\sqrt{N}}\right) \leqslant 2\exp(-t),$$

$$\mathbb{P}\left(\|\theta_N - \theta_\star\|^2 \geqslant \frac{15t}{R^2} + 6\|\theta_0 - \theta_\star\|^2\right) \leqslant 2\exp(-t).$$

Remarques : 1) Une autre méthode de démonstration serait d'utiliser l'inégalité de Freedman. 2) Les quantités θ_n et $\overline{\theta_n}$ ne convergent pas forcement vers $\theta_\star$. Par ailleurs, $\theta_\star$ n'est pas nécessairement unique.

6 Convergence des gradients

On va montrer que la norme du gradient au carré converge en $O(1/n)$ à l'aide de la propriété d'autoconcordance.

Proposition 6.1. *Sous les hypothèses **(A1)-(A7)**, avec un pas γ on a pour $p \geqslant 0$:*

$$\left(\mathbb{E}\|f'(\overline{\theta_n})\|^{2p}\right)^{1/2p} \leqslant \frac{R}{\sqrt{n}}\left[8\sqrt{p} + \frac{4p}{\sqrt{n}} + 40R^2\gamma p\sqrt{n} + \frac{3}{\gamma p\sqrt{n}}\|\theta_0 - \theta_\star\|^2 + \frac{3}{\gamma R\sqrt{n}}\|\theta_0 - \theta_\star\|\right].$$

Corollaire 6.2. *Sous les hypothèses **(A1)-(A7)**, avec un pas $\gamma = 1/(2R^2\sqrt{N})$ on a pour $p \geqslant 0$:*

$$\left(\mathbb{E}\|f'(\overline{\theta_N})\|^{2p}\right)^{1/2p} \leqslant \frac{R}{\sqrt{N}}\left[8\sqrt{p} + \frac{4p}{\sqrt{n}} + 20p + 6R^2\|\theta_0 - \theta_\star\|^2 + 6R\|\theta_0 - \theta_\star\|\right].$$

Remarques : 1) En prenant $p = 1$, on remarque que $\|f'(\overline{\theta_N})\|^2$ converge en $O(1/N)$.
2) Les normes L^p de $f'(\overline{\theta_n})$ sont majorées par une fonction affine en p, ce qui donne un comportement sous exponentiel.

7 Utilisation avancée de l'autoconcordance

On va utiliser un lemme technique :

Lemme 7.1. *Soit f une fonction convexe trois fois différentiable de $\mathcal{H}$ dans $\mathbb{R}$ qui vérifie la propriété d'autoconcordance **(A2)**. Soit $\theta_\star$ un minimum global de f et μ la plus petite valeur propre de $f''(\theta_\star)$, qui est supposée strictement positive. Si on a $\|f'(\theta)\|R/\mu \leqslant 3/4$ alors on a :*

$$\|\theta - \theta_\star\|^2 \leqslant \frac{4\|f'(\theta)\|^2}{\mu^2}, \quad f(\theta) - f(\theta_\star) \leqslant \frac{2\|f'(\theta)\|^2}{\mu}.$$

Théorème 7.2. *Sous les hypothèses **(A1)-(A7)**, avec un pas $\gamma = 1/(2R^2\sqrt{N})$ et en notant $\mu > 0$ la plus petite valeur propre de la Hessienne de f à l'unique minimum global $\theta_\star$, on a :*

$$\mathbb{E}f(\overline{\theta_N}) - f(\theta_\star) \leqslant \frac{R^2}{N\mu}\left(5R\|\theta_0 - \theta_\star\| + 15\right)^4,$$

$$\mathbb{E}\|\overline{\theta_N} - \theta_\star\|^2 \leqslant \frac{R^2}{N\mu^2}\left(6R\|\theta_0 - \theta_\star\| + 21\right)^4.$$

Démonstration. D'après la proposition 6.1 on a une majoration des moments de $f'(\overline{\theta_n})$, d'où un comportement sous exponentiel et donc un contrôle de $\|f'(\overline{\theta_n})\|$. On peut donc appliquer le lemme précédent 7.1 à $\overline{\theta_n}$. Reste à exprimer l'espérance comme une intégrale de la fonction de queue, faire du découpage, puis à utiliser des résultats de concentration sur les fonctions de queues. □

Remarque : On a pris $\gamma = 1/(2R^2\sqrt{N})$, il serait intéressant de prendre $\gamma_n = O(1/R^2\sqrt{n})$.
On aurait sans doute des termes en $\log n$ en plus.