Fiche

Charles Vin

Date

1 Formule et définition

1.1 Base

- Produit scalaire : $\langle x, y \rangle = x^T y = \sum x_i y_i$
- Norme : $\|x\| = \sqrt{\langle x, x \rangle}$
- Bilinéarité du produit scalaire :
  - $k \langle x, y \rangle = \langle kx, y \rangle = \langle x, ky \rangle$
  - $\langle z, x + y \rangle = \langle z, x \rangle + \langle x, y \rangle$
  - $\langle x, y \rangle + \langle x, z \rangle = \langle x, y + z \rangle$
- Dérivé norme : $\frac{d}{dt} \|f(t)\| = \frac{\langle f(t), f'(t) \rangle}{\|f(t)\|}$
- Dérivé norme au carré : $\frac{d}{dt} \|f(t)\|^2 = 2 \langle f(t), f'(t) \rangle$
- Identité remarquable : $\|a + b\|^2 = \|a\|^2 + \|b\|^2 + 2 \langle a, b \rangle$
- GD : $\theta_{t+1} = \theta_t - \gamma \nabla F(\theta_t)$
- Polyak-Ruppert averaging : $\bar{\theta}_T = \frac{1}{T} \sum_{t=1}^{T} \theta_t$
- KKT :

1.2 Inégalité

- Inégalité triangulaire : $\|x + y\| \leq \|x\| + \|y\|$
- Inégalité de Cauchy : $|\langle x, y \rangle| \leq \|x\| \|y\|$

1.3 Propriétés importantes

- Convexity props
  - Convexity : under chords : $F(\eta\theta + (1 - \eta)\theta') \leq \eta F(\theta) + (1 - \eta) F(\theta')$
  - Convexity : increasing slopes $\langle \nabla F(\theta) - \nabla F(\theta'), \theta - \theta' \rangle \geq 0$
  - Convexity + diff : $F(\theta') \geq F(\theta) + \langle \nabla F(\theta), \theta' - \theta \rangle \Leftrightarrow F(\theta') - F(\theta) \geq \langle \nabla F(\theta), \theta' - \theta \rangle$
  - Convexity + $\mathcal{C}^2$: Hessienne SDP $\forall \theta, Hess_F(\theta) \succeq 0$
- μ-strongly convex, $\mu > 0$.
  - μ-convexity : $F(\eta\theta + (1 - \eta)\theta') \leq \eta F(\theta) + (1 - \eta) F(\theta') - \frac{\eta(1-\eta)\mu}{2} \|\theta - \theta'\|_2^2$,
  - μ-convexity + diff : $F(\theta') \geq F(\theta) + < \nabla F(\theta), \theta' - \theta > + \frac{\mu}{2} \|\theta - \theta'\|_2^2$,
  - μ-convexity : $\langle \nabla F(\theta) - \nabla F(\theta'), \theta - \theta' \rangle \geq 0 + \mu \|\theta - \theta'\|_2^2$
  - μ-convexity + $\mathcal{C}^2$: Hessienne SDP $\forall \theta, Hess_F(\theta) \succeq \mu Id$ (SDP)
- k-lipschitzienne : $|f(x) - f(y)| \leq k |x - y|$ Bouger dans l'espace d'arriver fait bouger k fois plus dans l'espace de départ.
- L-Smooth : = gradient Lipschitz , $\|\nabla F(\theta) - \nabla F(\theta')\| \leq L \|\theta - \theta'\|$
- Co-coercivity = L-Smooth + Convexe :
  - $\frac{1}{L} \|\nabla F(\theta) - \nabla F(\theta')\|_2^2 \leq \langle \nabla F(\theta) - \nabla F(\theta'), \theta - \theta' \rangle \Leftrightarrow$
  - $\Leftrightarrow \|\nabla F(\theta_t)\|^2 \leq L \langle \nabla F(\theta_t), \theta_t - \theta^\star \rangle$

— Descent Lemma : = (Convexity + diff avec un $\leq$) + le terme L-Smooth

$$F(\theta') \leq F(\theta) + \langle \nabla F(\theta), \theta' - \theta \rangle + \frac{L}{2} \left\| \theta' - \theta \right\|^2 .$$

$$F(\theta') - F(\theta) \leq \langle \nabla F(\theta), \theta' - \theta \rangle + \frac{L}{2} \left\| \theta' - \theta \right\|^2 .$$

2 Technique de preuve

— Penser au $\pm$ pour faire apparaitre un terme voulu
— $\nabla F(\theta^{\infty}) \approx \nabla F(\theta^{\star}) = 0$
— Trick de l'intégrale, souvent suivie d'un $\pm \nabla F()$

$$\begin{aligned} F(x - \gamma y) - F(x) &= F(x - \gamma y) - F(\theta - 0 \times y) \\ &= [F(x - \tau y)]_0^{\gamma} \\ &= \int_0^{\gamma} \langle -y, \nabla F(x - \tau y) \rangle \, d\tau \end{aligned}$$

$$\begin{aligned} F(\theta_{t+1}) - F(\theta_t) &= F(\theta_t - \gamma \nabla F(\theta_t)) - F(\theta_t) \\ &= [F(\theta_t - \tau \nabla F(\theta_t))]_{\tau=0}^{\gamma} \\ &= - \int_0^{\gamma} \langle \nabla F(\theta_t), \nabla F(\theta_t - \tau \nabla F(\theta_t)) \rangle \, d\tau \end{aligned}$$

— Si on a des inégalités avec du θ_1 et des sommes, potentiel somme d'inégalités
— On utilise souvent la cocoercivity du gradient avec $\nabla F(\theta^{\star}) = 0$

$$\|\nabla F(\theta_t)\| \leq L \langle \nabla F(\theta_t), \theta_t - \theta^{\star} \rangle .$$

3 Gros gros plan du cours

— Basic of deterministic optim
  — GD when L-Smooth
  — GD when not L-Smooth
— SGD
  — Tourne autour de la solution ($Var(\nabla F_i(\theta^{\star})) = 1/3$)
  — Polyak averaging fix ça
  — Vanishing step size