Lecture 11 - landscape of rates

Tuesday, February 21, 2023 2:24 PM

today:
- finish convergence for SGD
- rate landscape
- CRF / structured SVM optimization

$$\underbrace{\mathbb{E}[\|x_{t+1}-\tilde{x}\|^2]}_{r_{t+1}} \leq (1-\mu\gamma_t)\underbrace{\mathbb{E}[\|x_t-\tilde{x}\|^2]}_{r_t} - 2\gamma_t\,[\underbrace{\mathbb{E}f(x_t)-f(\tilde{x})}_{\varepsilon_t}] + \gamma_t^2 B^2$$

⊛ non-strongly convex setting ($\mu = 0$)

set $\tilde{x}$ to be some minimizer x^* of f i.e. $f(x^*) = \min_{x\in C} f(x)$

[for better rate, let $\tilde{x}(x_0) \triangleq x^* = \operatorname{argmin}_{x\in X^*} \|x-x_0\|^2$]

let $r_t \triangleq \mathbb{E}\|x_t - x^*\|^2$

let $\boxed{\varepsilon_t \triangleq \mathbb{E}[f(x_t) - f(x^*)]}$ expected suboptimality error

$$r_{t+1} \leq r_t - 2\gamma_t\varepsilon_t + \gamma_t^2 B^2 \quad \forall t$$

$$2\gamma_t[\varepsilon_t] \leq r_t - r_{t+1} + \gamma_t^2 B^2 \quad \forall t$$

$$\Rightarrow 2\sum_{t=0}^{T}\gamma_t\varepsilon_t \leq \underbrace{r_0 - r_{T+1}}_{} + \left(\sum_{t=0}^{T}\gamma_t^2\right)B^2$$

telescoping sum $\sum_{t=0}^{T}(r_t - r_{t+1})$

$$2\left(\sum_t \gamma_t\right)\min_{0\leq t\leq T}\varepsilon_t \leq \quad \leq r_0 + \left(\sum_t \gamma_t^2\right)B^2$$

a) $\Rightarrow$
$$\boxed{\min_{0\leq t\leq T}\varepsilon_t \leq \frac{r_0 + \left(\sum_t \gamma_t^2\right)B^2}{2\left(\sum_t \gamma_t\right)}}$$

note: $\min_{0\leq t\leq T}\varepsilon_t \to 0$ when $\dfrac{\sum_t \gamma_t^2}{\sum_t \gamma_t} \xrightarrow{T\to\infty} 0$

use $\gamma_t^* = \dfrac{r_0}{B\sqrt{T+1}}$ to minimize RHS

$$\Rightarrow \boxed{\min_{t\leq T}\varepsilon_t \leq \frac{B r_0}{\sqrt{T+1}}}$$

⊛ [illegible] $\min \varepsilon_t \leq O(\log(T+1))$

$\left| \begin{matrix} t \leq T & \sqrt{T+1} \end{matrix} \right|$

(*) can also show that with $\gamma_t = \frac{A}{\sqrt{t+1}}$, $\min_{t \leq T} \varepsilon_t \leq O\left(\frac{\log(T+1)}{\sqrt{T+1}}\right)$

and if set C is bounded can show $O\left(\frac{\operatorname{diam}(C)}{\sqrt{T+1}}\right)$ rate

b) for $\hat{x}_T = \sum_t p_t x_t$

since f is convex $f(\hat{x}_T) = f\left(\sum_t p_t x_t\right) \overset{\text{Jensen's}}{\leq} \sum_t p_t f(x_t)$

$$\mathbb{E} f(\hat{x}_T) - f(x^*) \leq \sum_t p_t \underbrace{\left[\mathbb{E} f(x_t) - f(x^*)\right]}_{\varepsilon_t}$$

strongly convex case ($\mu > 0$)

$$r_{t+1} \leq (1 - \mu\gamma_t) r_t - \underbrace{2\gamma_t}_{\text{divide}} \varepsilon_t + \gamma_t^2 B^2$$

use $\boxed{\gamma_t = \frac{2}{\mu(t+2)}}$

$$\varepsilon_t \leq \frac{1}{2}(\gamma_t^{-1} - \mu) r_t - \frac{\gamma_t^{-1}}{2} r_{t+1} + \frac{\gamma_t}{2} B^2$$

$\hookrightarrow \frac{\mu(t+2)}{2}$

$\gamma_t^{-1} = \frac{\mu(t+2)}{2}$

multiply ineq. by $(t+1)$

$$(t+1)\varepsilon_t \leq \frac{1}{2}(t+1)\left(\frac{\mu t + \cancel{2\mu} - \cancel{2\mu}}{2}\right) r_t - \frac{\mu(t+1)(t+2)}{4} r_{t+1} + \underbrace{\frac{(t+1)}{\cancel{2}} \cdot \frac{\cancel{2}}{\mu(t+2)} B^2}_{\leq \frac{B^2}{\mu}}$$

$$< \frac{\mu}{4}\Big[\underbrace{t(t+1) r_t}_{\triangleq u_t} - \underbrace{(t+1)(t+2) r_{t+1}}_{u_{t+1}}\Big] + \frac{B^2}{\mu}$$

(sum ineq.) $\hookrightarrow$ telescoping sum? (trick !!)

$$\Rightarrow S_T \sum_{t=0}^{T} \frac{(t+1)}{S_T} \varepsilon_t \leq \frac{\mu}{4}\left[u_0 - u_{T+1}\right] + (T+1)\frac{B^2}{\mu}$$

let $p_t \triangleq \frac{t+1}{S_T}$ where $S_T \triangleq \sum_{t=0}^{T}(t+1) = \frac{(T+1)(T+2)}{2}$

$$S_T \sum_{t=0}^{T} p_t \varepsilon_t \leq \frac{\mu}{4}\left[0 - (T+1)(T+2) r_{T+1}\right] + (T+1)\frac{B^2}{\mu}$$

$$\sum_{t=0}^{T} p_t \varepsilon_t + \underbrace{\frac{\mu}{4}\left(\frac{(T+1)(T+2)}{S_T} r_{T+1}\right)}_{\frac{\mu}{2} r_{T+1}} \leq \underbrace{\frac{T+1}{S_T} \frac{B^2}{\mu}}_{\frac{2}{T+2}\frac{B^2}{\mu}} \quad (\ddagger)$$

$$\underbrace{\qquad\qquad}_{\frac{\mu}{2} r_{T+1}} \qquad \underbrace{\qquad}_{\frac{2}{T+2}\frac{B^2}{\mu}}$$

let $\hat{x}_T \triangleq \sum_{t=0} \rho_t x_t$ (weighted avg.)

$$f(\hat{x}_T) \leq \sum_t \rho_t f(x_t) \implies \mathbb{E} f(\hat{x}_T) - f(x^*) \leq \sum_t \rho_t \varepsilon_t \leq \frac{2B^2}{(T+2)\mu} \quad (\ddagger)$$

thus $\boxed{\mathbb{E} f(\hat{x}_T) - f(x^*) \leq \frac{2B^2}{\mu(T+2)}}$ vs. $O\left(\frac{1}{\sqrt{T}}\right)$ rate when $\mu = 0$

also $\boxed{\mathbb{E} \|x_{T+1} - x^*\|^2 \leq \frac{4}{\mu^2} \frac{B^2}{T+2}}$

15h24

Landscape of global convergence rates

f is convex: rate on suboptimality $f(x_t) - f(x^*) \leq$ ____

for stochastic setting $\mathbb{E} f(\hat{x}_t) - f(x^*) \leq$ ____

$r_0 \geq \text{dist}(x_0, X^*)$

assumptions	rate (deterministic/batch)	stochastic setting	finite sum $\frac{1}{n}\sum_i f_i(x)$ (special case)
1) non-smooth $\|\partial f\| \leq B$	$O\left(\frac{B r_0}{\sqrt{t}}\right)$ subgradient method	$O\left(\frac{B r_0}{\sqrt{t}}\right)$	
2) smooth L-Lipschitz ∇f	$O\left(\frac{L r_0^2}{t}\right)$ gradient method	$O\left(\frac{\square}{\sqrt{t}}\right)$ SGD	$O\left(\frac{\sqrt{n}\, L}{t}\right)$
	$O\left(\frac{L r_0^2}{t^2}\right)$ Nesterov method "optimal method"; matching lower bound		SAG/SAGA/SVRG ↳ "loopless version" [Hoffman & al.]
f is μ-strongly convex: 3) non-smooth $\|\partial f\| \leq B$	$O\left(\frac{B^2}{\mu t}\right)$ subg. method	$O\left(\frac{B^2}{\mu t}\right)$	
4) smooth L-Lipschitz	$O\left(\exp\left(-\frac{\mu}{L} t\right)\right)$ grad.	$O\left(\frac{\square}{\mu t}\right)$	$O\left(\exp\left(-\min\left\{\frac{1}{n}, \frac{\mu}{L}\right\} t\right)\right)$ SAG/SAGA/SVRG

L-Lipschitz | $O(\exp(-\frac{\mu}{L}t))$ gradient | $O(\frac{1}{\mu t})$ | $O(\exp(-\min(\frac{1}{n}, \frac{\mu}{L})t))$

$O(\exp(-\sqrt{\frac{\mu}{L}}t))$ Nesterov "optimal"

SAG/SAGA/SVRG

faster → Katyusha alg.

interpolation regime $\mathbb{E}_\xi \|\nabla_x h(x^*, \xi)\|^2 = 0$

↳ overparameterized regime

⟶ get faster rate for SGD

(*) note: projecting gives same rates

more generally, proximal gradient method as well

setup: "composite smooth opt." $\min_x \underbrace{f(x)}_{\text{smooth}} + \underbrace{h(x)}_{\text{non-smooth but "simple"}}$

constrained opt. is special case: $h(x) \triangleq \delta_C(x) \triangleq \begin{cases} +\infty & \text{if } x \notin C \\ 0 & \text{o.w.} \end{cases}$

proximal gradient method:

prox. gradient step.

$$x_{t+1} \triangleq \arg\min_x \underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \nu \frac{L}{2} \|x - x_t\|^2}_{\frac{\nu L}{2} \|x - (x_t - \frac{1}{\nu L} \nabla f(x_t))\|_2^2 + cst.} + h(x)$$

* if $h = \delta_C \implies$ projected gradient method

* but can also run on other "simple" h e.g. $h(x) = \|x\|_1$ (Lasso type problem)

⟶ prox step becomes "soft-thresholding" operator

→ get same rate as unconstrained

[accelerated prox. gradient for ℓ_1 = FISTA ; SOTA for deterministic ℓ_1-reg. problem (small n)]

optimization of $\hat{f}(w)$

$$\hat{f}(w) = R(w) + \frac{1}{n} \sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}; w)$$

say $R(w) = \frac{1}{2}\|w\|_2^2$

recall $h_w(x) = \arg\max_{\tilde{y} \in \mathcal{Y}} \langle w, \varphi(x, \tilde{y}) \rangle$

CRF:

$$f_{CRF}(x,y;w) \triangleq \log\left(\sum_{\tilde{y}} \exp(\langle w, \varphi(x,\tilde{y})\rangle)\right) - \langle w, \varphi(x^{(i)}, y^{(i)})\rangle$$ neg. cond. likelihood loss

here $\hat{f}_{CRF}(w)$ is L-smooth & λ-strongly convex $\quad p_w(\tilde{y}|x) \propto \exp(s(\tilde{y}))$

weighted avg. SGD $\to$ get a rate of $O\left(\frac{B}{\mu t}\right)$

what do we need to run SGD

$$\nabla_w f(x,y;w) = \underbrace{\frac{1}{\sum_{\tilde{y}} \exp(s(\tilde{y}))} \sum_{\tilde{y}} \exp(s(\tilde{y}))}_{\to\, p_w(\tilde{y}|x)} \left[\varphi(x,\tilde{y}) - \varphi(x^{(i)}, y^{(i)})\right]$$

$$= \mathbb{E}_{\tilde{y}|x;w}\left[\varphi(x,\tilde{y}) - \varphi(x,y)\right]$$

CRF: $\varphi(x,\tilde{y}) = \sum_{c \in \mathcal{C}} \varphi_c(x, \tilde{y}_c)$

then $\mathbb{E}_{\tilde{y}|x}[\varphi(x,\tilde{y})] = \sum_{c\in\mathcal{C}} \underbrace{\mathbb{E}_{\tilde{y}_c|x}}_{\text{marginal over } \tilde{y}_c}[\varphi_c(x,\tilde{y}_c)]$

use sum-product alg on trees e.g.
or junction tree alg for small treewidth graph

"inference oracle"

structured SVM

$$f_{hinge}(x^{(i)}, y^{(i)}; w) = \max_{\tilde{y}\in\mathcal{Y}} \langle w, \varphi(x^{(i)},\tilde{y})\rangle + \ell(y^{(i)},\tilde{y}) - \langle w, \varphi(x^{(i)}, y^{(i)})\rangle$$

let $\ell_i(\tilde{y}) \triangleq \ell(y^{(i)}, \tilde{y})$

$H_i(w) \triangleq f_{hinge}(x^{(i)}, y^{(i)}, w) \;\to\; = \max_{\tilde{y}} \ell_i(\tilde{y}) - \underbrace{\langle w, \psi_i(\tilde{y})\rangle}_{m(\tilde{y})}$ "margin element"

$\psi_i(\tilde{y}) \triangleq \varphi(x^{(i)}, y^{(i)}) - \varphi(x^{(i)}, \tilde{y})$

note: if $\langle w, \psi_i(\tilde{y})\rangle > 0 \;\; \forall \tilde{y} \neq y^{(i)}$
then $h_w(x^{(i)}) = y^{(i)}$

$H_i(w;\tilde{y}) \triangleq \ell_i(\tilde{y}) - \langle w, \psi_i(\tilde{y})\rangle$

so $H_i(w) = \max_{\tilde{y}} H_i(w;\tilde{y})$

structured SVM objective (non-smooth unconstrained form)

$$\boxed{\min_w \frac{\lambda \|w\|^2}{2} + \frac{1}{n}\sum_{i=1}^n H_i(w)}$$

unconstrained form) $\vdash$

(*) this fits stochastic sub. method framework : $f(w) = \mathbb{E}_i\, h(w,i)$

where $h(w,i) \triangleq \frac{\lambda}{2}\|w\|^2 + H_i(w)$

now a subgradient of $h(w,i)$

$$h'(w,i) = \lambda w - \psi_i(\hat{y}_i(w))$$

$\hookrightarrow \underset{\tilde{y}\in\mathcal{Y}_i}{\arg\max}\; l_i(\tilde{y}) - \langle w, \psi_i(\tilde{y})\rangle$

"loss-augmented inference"

$$\mathbb{E}_i\, h'(w,i) = \lambda w - \frac{1}{n}\sum_{i=1}^{n} \psi_i(\hat{y}_i(w)) = f'(w)$$

[batch subgradient]

<u>convergence rate</u>:

have f is λ-strongly convex

suppose that $\|\psi_i(\tilde{y})\|_2 \le R \quad \forall i \;\&\; \tilde{y} \in \mathcal{Y}_i$

then one can show that if $\gamma_t = \frac{2}{\lambda(t+2)}$, then $\mathbb{E}\|g_t\|^2 \le 4R^2 \leftarrow$ gives our B^2

<u>and</u> $w_0 = 0$

$\longrightarrow O\left(\frac{R^2}{\lambda t}\right)$

[exercise: adapt of App. A of arXiv note Lacoste-Julien & al. 2012]