Lecture 11 - landscape of rates

Tuesday, February 11, 2020 14:31

today: · finish proof for SGD
· rate landscape
· CRF / structured SVM optimization

$$\mathbb{E}[\|x_{t+1}-\tilde{x}\|^2] \le (1-\mu\gamma_t)\,\mathbb{E}[\|x_t-\tilde{x}\|^2] - 2\gamma_t\left[\mathbb{E}f(x_t)-f(\tilde{x})\right] + \gamma_t^2 B^2$$

(true even if $\mu=0$; for γ_t small enough, we have $\mathbb{E}\|x_t-\tilde{x}\|^2$ decreases with t for any $\tilde{x}\in C$ s.t. $f(\tilde{x}) \le \mathbb{E}f(x_t)$)

$r(t)$

we have $r(t+1) \le r(t)$

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

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

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

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

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

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

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

$$\Rightarrow 2\sum_{t=0}^{T}\gamma_t\varepsilon_t \le \underbrace{r_0 - r_{T+1}}_{\text{telescoping sum } \sum_{t=0}^{T}(r_t - r_{t+1})} + \left(\sum_t \gamma_t^2\right)B^2$$

$$2\left(\sum_t \gamma_t\right)\min_{0\le t\le T}\varepsilon_t \le 2\sum_t \gamma_t\varepsilon_t \le \quad \| \quad \|$$

a) $\Rightarrow$

$$\min_{0\le t\le T}\varepsilon_t \le \frac{r_0 + \left(\sum_{t=0}^{T}\gamma_t^2\right)B^2}{2\sum_{t=0}^{T}\gamma_t}$$

note: $\min \varepsilon_t \to 0$ when $\frac{\sum_{t=0}^{T}\gamma_t^2}{\sum_{t=0}^{T}\gamma_t} \xrightarrow{T\to\infty} 0$

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

$$\Rightarrow \min_{0\le t\le T}\varepsilon_t \le \frac{B r_0}{\sqrt{T+1}}$$

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

Jensen

since f is convex, $f(\hat{x}_T) = f\left(\sum_t p_t x_t\right) \le \sum_t p_t f(x_t)$

can also show that with $\gamma_t = \frac{A}{\sqrt{t+1}}$, $\min_{0 \le t \le T} \varepsilon_t \le O\left(\frac{\log(T+1)}{\sqrt{T+1}}\right)$

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

strongly convex case ($\mu > 0$)

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

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

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

multiply ineq. by $(t+1)$

$(t+1)\varepsilon_t \le \frac{1}{2}(t+1)\left(\frac{\mu t + 2\mu - 2}{2}\right) r_t - \frac{\mu}{4}(t+1)(t+2) r_{t+1} + \frac{(t+1)}{2\mu(t+2)} B^2$

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

(sum ineq.) $\Rightarrow$ telescoping sum

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

($u_0 = 0$)

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

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

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

let $\hat{x}_T \triangleq \sum_{t=0}^{T} p_t x_t$ (weighted average)

$f(\hat{x}_T) \overset{\text{convexity}}{\le} \sum p_t f(x_t)$ $\quad \varepsilon_t$ ...

$f(\bar{x}_T) \overset{\text{convexity}}{\leq} \sum_t p_t f(x_t)$

$\Rightarrow \mathbb{E}[f(\bar{x}_T) - f(x^*)] \leq \sum_t p_t \underbrace{\mathbb{E}[f(x_t) - f(x^*)]}_{\varepsilon_t} \overset{(*)}{\leq} \frac{2B^2}{T+2}$

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

Landscape of global convergence rates

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

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 special case $\frac{1}{n}\sum_{i=1}^n f_i(x)$	
1) non-smooth $\|\partial f\| \leq B$	$O\left(\frac{Br_0}{\sqrt{t}}\right)$	subgradient method	$O\left(\frac{Br_0}{\sqrt{t}}\right)$		
2) smooth L-Lipschitz ∇f	$O\left(\frac{Lr_0^2}{t}\right)$	gradient method	$O\left(\frac{\square}{\sqrt{t}}\right)$ SGD	$O\left(\frac{\sqrt{n}L}{t}\right)$	SAG/SAGA SVRG & "dropless version" [Hoffman & al.]
	$O\left(\frac{Lr_0^2}{t^2}\right)$ (matching lower bound)	Nesterov method "optimal method"			
f is μ-strongly convex 3) non-smooth $\|\partial f\| \leq B$	$O\left(\frac{B^2}{\mu t}\right)$	subgradient meth.	$O\left(\frac{B^2}{\mu t}\right)$		
4) smooth L-Lipschitz	$O\left(\exp\left(-\frac{\mu}{L}t\right)\right)$	gradient method	$O\left(\frac{\square}{\mu t}\right)$	$O\left(\exp\left(-\min\left\{\frac{1}{n}, \frac{\mu}{L}\right\}t\right)\right)$	SAG/SAGA/SVRG
	$O\left(\exp\left(-\sqrt{\frac{\mu}{L}}t\right)\right)$	Nesterov "optimal"			

15h55

(*) note: projecting gives the same rates

more generally, proximal gradient method as well smooth non-smooth

setup: "composite smooth opti." $\min_x f(x) + h(x)$

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

proximal gradient method:

$$x_{t+1} = \underset{x}{\operatorname{argmin}} \; f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta} \|x - x_t\|^2 + h(x)$$

(prox step)

$$\frac{\eta L}{2} \left\| x - \left(x_t - \frac{1}{\eta L} \nabla f(x_t) \right) \right\|^2 + \text{cst.}$$

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

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

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

Optimization of $\hat{\ell}(w)$

$$\hat{\ell}(w) = R(w) + \frac{1}{n} \sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}; w)$$ say $R(w) = \frac{\lambda}{2}\|w\|_2^2$

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

CRF:

$$\ell_{CRF}(\tau, y; w) \triangleq \log \left(\sum_{\tilde{y}} \exp(\langle \tau w, \varphi(x, \tilde{y}) \rangle) \right) - \langle w, \varphi(x^{(i)}, y^{(i)}) \rangle$$

← reg. log-likelihood loss (conditional)

here $\hat{\ell}_{CRF}(w)$ is L-smooth & λ-strongly convex

$p_w(\tilde{y}|x) \propto \exp(s_{\tilde{y}})$

weighted avg. SGD → get a rate of $O\left(\frac{\square}{\text{nb}}\right)$

What do we need to run SGD?

compute $\nabla_w \ell(x, y; w) = \frac{1}{\sum_{y'} \exp(s_{y'})} \sum_{\tilde{y}} \exp(s_{\tilde{y}}) \varphi(x, \tilde{y}) - \varphi(x, y)$

$\frac{\exp(s_{\tilde{y}})}{\sum_{y'} \exp(s_{y'})} = p_w(\tilde{y}|x)$

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

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

, marginal over $\tilde{y}_c$,

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

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

→ use sum-product alg on trees eg or junction tree alg for small tree width graphs

structured SVM:

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

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

$H_i(w) \triangleq \ell_{\text{hinge}}(x^{(i)}, y^{(i)}; w) \longrightarrow = \max_{\tilde{y} \in \mathcal{Y}} \ell_i(\tilde{y}) - \langle w, \Psi_i(\tilde{y}) \rangle$

$\Psi_i(\tilde{y}) \triangleq \varphi(x^{(i)}, y^{(i)}) - \varphi(x^{(i)}, \tilde{y})$ $\quad m(\tilde{y})$ "margin element"

$H_i(w; \tilde{y}) \triangleq \ell_i(\tilde{y}) - \langle w, \Psi_i(\tilde{y}) \rangle$ note: if $\langle w, \Psi_i(\tilde{y}) \rangle > 0 \ \forall \tilde{y} \neq y^{(i)}$

then $h_w(x^{(i)}) = y^{(i)}$

structured SVM objective (nonsmooth unconstrained form)

$$\min_w \ \underbrace{\frac{\lambda}{2}\|w\|^2}_{\text{strongly convex}} + \frac{1}{n}\sum_{i=1}^n H_i(w)$$

(*) this fits stochastic sub. method framework: $f(w) = \mathbb{E}_i[h(w,i)]$ where $h(w,i) \triangleq \underbrace{\frac{\lambda}{2}\|w\|^2}_{\text{strongly convex}} + H_i(w)$

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

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

$\hat{y}_i(w) \triangleq \operatorname{argmax}_{\tilde{y} \in \mathcal{Y}} \ell_i(\tilde{y}) - \langle w, \Psi_i(\tilde{y}) \rangle$ loss-augmented inference

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

convergence rate:

have f is λ-strongly convex

suppose that $\|\Psi_i(\tilde{y})\| \leq R \ \forall i, \tilde{y} \in \mathcal{Y}_i$

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

$\|\varphi(x,y)\|_2 \leq \sum_{c \in \mathcal{C}} \|\varphi_c(x, y_c)\|_2$

then one can show that with $\gamma_t = \frac{2}{\lambda(t+2)}$ and $w_0 = 0$, then $\mathbb{E}\|g_t\|^2 \leq 4R^2$ ← gives an R^2

[exercise: adapt App. A of Arxiv note Lacoste-Julien et al. 2012]

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