Lecture 11 - landscape of rates

Thursday, February 18, 2021 13:36

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

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

(*) 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 brute, let $\tilde{x}(x_0) = x^* = \operatorname{argmin}_{x \in X^*} \|x - x_0\|^2$]

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

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

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

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

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

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

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

a) $\Rightarrow$

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

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

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

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

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

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

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

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

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

strongly convex case ($\mu > 0$)

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

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

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

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

multiply ineq. by $(t+1)$

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

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

telescoping sum! (trick)

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

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

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

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

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

Convexity ε_t $(\dagger)$

Let $\hat{x}_T \triangleq \sum_{t=2}^{T} p_t x_t$ (weighted avg.)

$f(\hat{x}_T) \leq \sum p_t f(x_t)$ by convexity $\Rightarrow$

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

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

14h23

Landscape of global convergence rates

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

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

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

assumptions	rate (deterministic/batch)		stochastic setting	finite sum $\frac{1}{n}\sum_{i=1}^{n} f_i(x)$ special case
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 ↳ "loopless" version [Hofman's Sol.]
	$O\left(\frac{Lr_0^2}{t^2}\right)$ matching lower bound	Nesterov method "optimal method"		
f is μ-strongly convex 3) non-smooth $\|\partial f\| \leq B$	$O\left(\frac{B^2}{\mu t}\right)$	subg. method	$O\left(\frac{B^2}{\mu t}\right)$	
4) smooth L-Lipsch	$O(\exp(-\frac{\mu}{L}t))$	grad.	$O\left(\frac{\square}{\mu t}\right)$	$O(\exp(-\min\{\frac{1}{n}, \frac{\mu}{L}\}t))$ SAG/SAGA/SVRG
	$O(\exp(-\sqrt{\frac{\mu}{L}}t))$	Nesterov "optimal"		

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

↳ overparameterized regime $\Rightarrow$ get faster rates for SGD

(*) note: projecting gives the same rates

more generally, proximal gradient method as well

setup: "composite smooth optim." $\min_x f(x) + h(x)$ (smooth + non-smooth)

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:

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

(prox. step)

$$\frac{\nu L}{2} \left\| x - \left(x_t - \frac{1}{\nu L} \nabla f(x_t) \right) \right\|_2^2 + \text{cst.}$$

* 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)

→ prox step becomes "soft-thresholding" operator

→ get same rate convergence as unconstrained

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

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

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

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

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

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. log likelihood loss

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

$p_w(\tilde{y}|x)$ or $\exp(s(\tilde{y}))$

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

what do we need to run SGD?

$$\nabla_w f(x, y; w) = \frac{1}{\sum_{\tilde{y}} \exp(s(\tilde{y}))} \sum_{\tilde{y}} \exp(s(\tilde{y})) \left[\varphi(x, \tilde{y}) \right] - \varphi(x, y)$$

$\to p_w(\tilde{y}|x)$

to run SGD?

$\sum_{\tilde{y}} \exp(s(\tilde{y}))$ → $p_w(\tilde{y}|x)$

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

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

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

$\hat{y}_c$ marginal over $\hat{y}$

use sum-product alg on trees CG or junction tree alg. for small tree width graph

structured SVM

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

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

$H_i(w) \triangleq \text{hinge}(x^{(i)}, y^{(i)}; w)$

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

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

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

$= \max_{\tilde{y} \in \mathcal{Y}} \ell_i(\tilde{y}) - \langle w, \Psi_i(\tilde{y}) \rangle$

$m(\tilde{y})$ "margin element"

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 (non-smooth unconstrained form)

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

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

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

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

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

$\hat{y}_i(w) = \text{argmax}_{\tilde{y} \in \mathcal{Y}} \ell_i(\tilde{y}) - \langle w, \Psi_i(\tilde{y}) \rangle$

loss-augmented "inference"

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

[batch subgradient]

convergence rate:

convergence rate:

here f is λ-strongly convex

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

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 our B^2

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

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

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

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