Lecture 11 - landscape of rates

Tuesday, February 20, 2024 9:28 AM

today: • finish for SGD
• rate landscape
• CRF / structured sum optimization

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

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

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

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

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

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

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

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

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

note: $\sum_{t=0}^{T}\gamma_t \xrightarrow{T\to\infty} \infty$

if $\dfrac{\sum_t \gamma_t^2}{\sum_t \gamma_t} \xrightarrow{T\to\infty} 0$

then $\min_{t\le T}\varepsilon_t \to 0$

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

($r_0 = \|x_0 - x^*\|^2$)

$$\Rightarrow \quad \min_{t\le T}\varepsilon_t \le \frac{Br_0}{\sqrt{T+1}}$$

b) Def $\hat{x}_T = \sum_t \rho_t x_t$

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

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

(b) 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 ($\mu > 0$)

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

divide

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

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

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

multiply ineq. by $(t+1)$

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

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

telescoping sum? (trick)

(sum ineq.)

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

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

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

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

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

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

$$\mathbb{E} f(\hat{x}_T) - f^* \le \sum_t p_t \varepsilon_t \overset{(\ddagger)}{\le} \frac{2B^2}{(T+2)\mu}$$

by convexity

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

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

$$\mathbb{E}\varepsilon_{T+1} \le \frac{L}{2}\mathbb{E}\|x_{T+1} - x^*\|^2 \longrightarrow \frac{2}{\mu}\left(\frac{L}{\mu}\right)\frac{B^2}{(T+2)}$$

Ch24

Landscape of global convergence rates

f is convex rate on suboptimality $f(x_t) - f(x^*) \le$ ~~~

for stochastic setting: $\mathbb{E} f(\hat{x}_t) - f(x^*) \le$ ~~~

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

assumptions	rate (deterministic / full batch)		stochastic setting	finite sum special case $\frac{1}{n}\sum_{i=1}^{n} f_i(x)$
1) non-smooth $\|\partial f\| \le B$	$O\left(\frac{Br_0}{\sqrt{t}}\right)$	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 ✗ "Logloss" [Hoffman's 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 $\|\partial f\| \le B$	$O\left(\frac{B^2}{\mu t}\right)$	subgradient method	$O\left(\frac{B^2}{\mu t}\right)$	
4) smooth L-Lipschitz	$O(\exp(\frac{\mu}{L}t))$ grad.		$O\left(\frac{\square}{\mu t}\right)$	$O(\exp(-\min\{\frac{1}{n}, \frac{\mu}{L}\}t))$ SAG/SAGA/SVRG
	$O(\exp(-\sqrt{\mu}\,t))$ Nesterov			

L-Lipschitz | $O(\exp(-\frac{\mu}{L}t))$ grad. | $O(\frac{1}{\mu t})$ | $O(\exp(-\min(\frac{1}{n}, \frac{\mu}{L})t))$
$O(\exp(-\sqrt{\frac{\mu}{L}}t))$ Nesterov "optimal" | SAG/SAGA/SVRG

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

↳ overparameterized regime

⇒ get faster rates for SGD

(*) note: projecting gives the same rates

more generally, proximal gradient method as well

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

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

* if $h = \delta_C$ ⇒ 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 no h (unconstrained or w/ $\|\cdot\|_2$ etc..)

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

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

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

recall $h_w(x) = \arg\max \langle w, \varphi(x, y) \rangle$

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

CRF: $\ell_{CRF}(x, y; w) \triangleq \log\left(\sum_{\tilde{y}} \exp(\underbrace{\langle w, \varphi(x, \tilde{y}) \rangle}_{\triangleq s(\tilde{y})}) \right) - \langle w, \varphi(x, y) \rangle$ ← neg. conditional likelihood loss

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

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

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

what do we need to run SGD?

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

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

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

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

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 $\tilde{y}_c$

"marginalization oracle"

→ use sum-product alg. on trees e.g.

or junction tree alg. for small treewidth graphs

Structured SVM

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

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

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

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

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

"hinge loss elements"

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

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

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

$$\min_w \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{n} \sum_{i=1}^{n} H_i(w)$$

this fits stochastic subgradient method framework: $\ell(w) = \mathbb{E}_i[\ell_i(w)]$

unconstrained form)

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

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

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

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

$$\hat{y}_i(w) \triangleq \operatorname*{argmax}_{\bar{y} \in \mathcal{Y}} \ell_i(\bar{y}) - \langle w, \psi_i(\bar{y}) \rangle$$

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

here f is λ-strongly convex

suppose that $\|\psi_i(\bar{y})\| \le R \quad \forall i, \bar{y} \in \mathcal{Y}$

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

and $w_0 = 0$

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

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