Lecture 11 - landscape of rates

Tuesday, February 12, 2019 14:29

today:
- rate landscape
- CRF / structured SVM optimization

Landscape of global convergence rates

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

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

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

assumptions	rate deterministic (batch)		stochastic setting		finite sum special case $\frac{1}{n}\sum_{i=1}^{n} f_i(x)$	
1) non-smooth, bounded subgradient $\lVert \partial f \rVert \leq B$	$O\left(\frac{B r_0}{\sqrt{t}}\right)$	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	$O\left(\frac{\sqrt{n} L}{t}\right)$	SAG/SAGA, SVRG → Hoffman's style
	$O\left(\frac{L r_0^2}{t^2}\right)$ (lower bound)	Nesterov method "optimal method"				
f is μ-strongly convex: 3) non-smooth $\lVert \partial f \rVert \leq B$	$O\left(\frac{B^2}{\mu t}\right)$	subgradient method	$O\left(\frac{B^2}{\mu t}\right)$			
f is μ-strongly convex: 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
	$O\left(\exp\left(-\sqrt{\frac{\mu}{L}}\, t\right)\right)$	Nesterov				

L-Lipschitz $\quad O(\exp(-\sqrt{\frac{\mu}{L}}t))$ Nesterov (lower bound) $\quad \mu t$ $\quad$ SAGA

(*) note: projecting gives the same rates

more generally, proximal gradient method as well

→ setup: "composite smooth optimization"

$$\min_x \underbrace{f(x)}_{\text{smooth}} + \underbrace{h(x)}_{\text{non-smooth}}$$

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:

prox step

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

if $h = \delta_C$

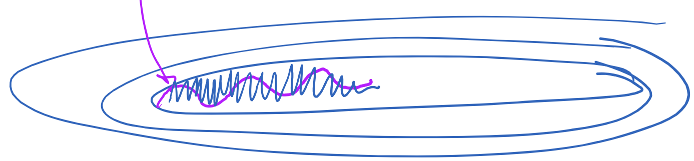
⇒ projected gradient method
= prox grad. method when $h(x) = \delta_C(x)$

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

(Lasso type problem)

[accelerated prox gradient = FISTA; state-of-the-art for deterministic ℓ_1-reg. problems]

Aside: Nesterov method behavior



15h16

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

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

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

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

CRF: $\ell_{CRF}(x, y; w) \triangleq \log\left(\sum_{\tilde{y}} \exp(\langle w, \varphi(x, \tilde{y}) \rangle)\right) - \langle w, \varphi(x^{(i)}, y^{(i)}) \rangle$ ← negative conditional log-likelihood loss

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

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

weighted average SGD → get a rate of $O\left(\frac{L}{\mu t}\right)$

what do we need? compute $\nabla_w \ell_{CRF}(x, y; w) = \sum_{\tilde{y}} \underbrace{\frac{\exp(s(\tilde{y}))}{\sum_{\tilde{y}'} \exp(s(\tilde{y}'))}}_{P_w(\tilde{y}|x)} \varphi(x, \tilde{y}) - \varphi(x, y)$

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

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 <u>sum-product</u> alg on trees e.g. or <u>junction tree</u> e.g. on small treewidth graphs

<u>structured SVM</u>:

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

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

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

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

$\hookrightarrow \langle w, \psi_i(\tilde{y})\rangle = m(\tilde{y})$ "margin"

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

note: if $\langle w, \psi_i(\tilde{y})\rangle > 0 \quad \forall \tilde{y} \neq y^{(i)}$

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

structured SVM objective (non-smooth unconstrained form)

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

(*) this fits stochastic subgradient 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 $h(w,i)$: $h'(w,i) = \lambda w - \psi_i(\hat{y}_i(w))$

$\hat{y}_i(w) \triangleq \arg\max\, \ell_i(\tilde{y}) - \langle w, \psi_i(\tilde{y})\rangle$

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

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

loss-augmented inference

[sidenote: soon, we will see that

$$w^* = \frac{1}{\lambda n}\sum_{i=1}^{n}\sum_{\tilde{y}\in\mathcal{Y}_i} \alpha_i^*(\tilde{y})\, \psi_i(\tilde{y})$$

$\sum_{\tilde{y}} \alpha_i(\tilde{y}) = 1$, $\alpha_i(\tilde{y}) \geq 0$

$\leftarrow$ true both for CRF & structured SVM]

for SVM
$$\begin{cases} \alpha_i^*(\tilde{y}) = 0 & \text{if } \; H_i(w^*;\tilde{y}) < \max_{\tilde{y}'} H_i(w^*;\tilde{y}') = H_i(w^*) \\ \alpha_i^*(\tilde{y}) > 0 & \text{"support vectors"} \end{cases}$$

<u>Convergence rate</u>:

here f is λ-strongly convex

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

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

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

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

[exercise: adapt appendix A of Arxiv note Lacoste-Julien & al. 2012]

$\rightarrow O\left(\frac{R^2}{\lambda t}\right)$ rate other approaches to optimize SVMstruct:

(UP) unconstrained primal

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

[unconstrained non-smooth]

(PQP) primal QP ($\to$ quadratic program)

$$\min_{w,(\xi_i)_{i=1}^n} \frac{\lambda \|w\|^2}{2} + \frac{1}{n}\sum_{i=1}^n \xi_i$$

$$\xi_i \geq H_i(w;\tilde{y}) \quad \forall \tilde{y} \in \mathcal{Y}_i \;\&\; \forall i$$

[constrained formulation

(smooth) convex QP with exp. # of linear constraints

1) generic approach to use convexity of loss-augmented decoding: [Taskar & al. ICML 2005]

idea: here, we suppose that loss-augmented decoding can be expressed as a "compact" maximization problem of a concave fct.

i.e. $$H_i(w) = \max_{\tilde{y} \in \mathcal{Y}} \ell_i(\tilde{y}) - \langle w, \psi_i(\tilde{y})\rangle = \max_{z \in Z} g_i(w;z)$$

$\tilde{y}$: discrete; z: cts.; Z: convex

where g_i is concave in z and convex in w

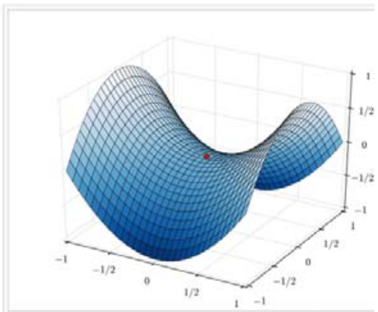
Z: • should not depend on w
• should have a tractable size

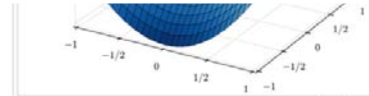
a) saddle point formulation:

$$\min_w \max_{z_i \in Z_i} \frac{\lambda \|w\|^2}{2} + \frac{1}{n}\sum_{i=1}^n g_i(w;z_i)$$

$$\min_w \max_z \mathcal{L}(w,z)$$

convex-concave saddle point problem




A saddle point (in red) on the graph of $z=x^2-y^2$ (hyperbolic paraboloid)

$$\min_w \max_z f(w,z)$$ convex-concave saddle point problem

(under reg. conditions min-max = max-min → "saddle point")

$$\forall z \quad f(w^*,z) \leq f(w^*,z^*) \leq f(w,z^*) \quad \forall w$$

in general

$$\min_w \max_z \geq \max_z \min_w$$

circular dependence

$$\begin{cases} w^* \in \arg\min_w f(w,z^*) \\ z^* \in \arg\max_z f(w^*,z) \end{cases}$$

⇒ it might not exist?

standard algorithms

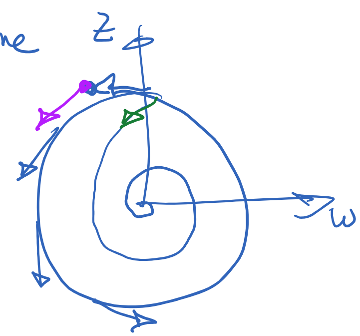
<u>extragradient algorithm:</u>

"lookahead step"

$$\begin{pmatrix} \tilde{w}_{t+1} \\ \tilde{z}_{t+1} \end{pmatrix} = \begin{pmatrix} w_t \\ z_t \end{pmatrix} + \gamma_t \begin{pmatrix} -\nabla_w f(w_t, z_t) \\ \nabla_z f(w_t, z_t) \end{pmatrix}$$

$$\begin{pmatrix} w_{t+1} \\ z_{t+1} \end{pmatrix} = \begin{pmatrix} w_t \\ z_t \end{pmatrix} + \gamma_t \begin{pmatrix} -\nabla_w f(\tilde{w}_{t+1}, \tilde{z}_{t+1}) \\ \nabla_z f(\text{''} \quad \text{''}) \end{pmatrix}$$

→ converges $O\left(\frac{1}{t}\right)$ for convex-concave game



$\langle w, z \rangle$

$$\begin{pmatrix} -z \\ w \end{pmatrix}$$

→ applied to structured SVM
[Taskar & al. JMLR 2006]