Convergence analysis

Convex and Lipschitz problems

$$\begin{array}{ll} \text{minimize}_{\boldsymbol{x}} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathcal{C} \end{array}$$

s.t.

- f is convex and Lipschitz continuous with respect to
  - φ is ρ-strongly convex w.r.t. a certain norm $\|\cdot\|$
  - $\|\boldsymbol{g}\|_* \le L_f$ for any subgradient $\boldsymbol{g} \in \partial f(\boldsymbol{x})$ at any point $\boldsymbol{x}$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$

$$\|g\|_* = \sup\{\langle g, x\rangle \mid \|x\| \le 1\}$$

Dual Norm

Convergence analysis

Theorem 5.3

Suppose f is convex and Lipschitz continuous (in the sense that $\|\boldsymbol{g}\|_* \leq L_f$ for any subgradient $\boldsymbol{g}$ of f) on $\mathcal{C}$. Suppose φ is ρ-strongly convex w.r.t. $\|\cdot\|$. Then

$$f^{\mathsf{best},t} - f^{\mathsf{opt}} \leq \frac{\sup_{\boldsymbol{x}\in\mathcal{C}} D_\varphi(\boldsymbol{x}, \boldsymbol{x}^0) + \frac{L_f^2}{2\rho}\sum_{k=0}^t \eta_k^2}{\sum_{k=0}^t \eta_k}$$

- If $\eta_t = \frac{\sqrt{2\rho R}}{L_f}\frac{1}{\sqrt{t}}$ with $R := \sup_{\boldsymbol{x}\in\mathcal{C}} D_\varphi(\boldsymbol{x}, \boldsymbol{x}^0)$, then

$$f^{\mathsf{best},t} - f^{\mathsf{opt}} \leq O\left(\frac{L_f\sqrt{R}}{\sqrt{\rho}}\frac{\log t}{\sqrt{t}}\right)$$

  - one can further remove the $\log t$ factor

Example: optimization over probability simplex

Suppose $\mathcal{C} = \Delta$ is the probability simplex, and pick $\boldsymbol{x}^0 = n^{-1}\mathbf{1}$

(1) set $\varphi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$, which is 1-strongly convex w.r.t. $\|\cdot\|_2$. Then

$$\sup_{\boldsymbol{x}\in\Delta} D_\varphi(\boldsymbol{x}, \boldsymbol{x}^0) = \sup_{\boldsymbol{x}\in\Delta} \frac{1}{2}\|\boldsymbol{x} - n^{-1}\mathbf{1}\|_2^2 = \sup_{\boldsymbol{x}\in\Delta} \frac{1}{2}\Big(\|\boldsymbol{x}\|_2^2 - \frac{1}{n}\Big) \le \frac{1}{2}$$

Then Theorem 5.3 says

$$\|x\|^2 - 2\langle x, \frac{\vec{1}}{n}\rangle + \|\frac{\vec{1}}{n}\|^2$$

$$f^{\mathsf{best},t} - f^{\mathsf{opt}} \le O\left(L_{f,2}\frac{\log t}{\sqrt{t}}\right)$$

$$= \|x\|_2^2 - \frac{2}{n} + \frac{1}{n}$$

if any subgradient $\boldsymbol{g}$ obeys $\|\boldsymbol{g}\|_2 \le L_{f,2}$

$$= \|x\|_2^2 - \frac{1}{n}$$

Example: optimization over probability simplex

Suppose $\mathcal{C} = \Delta$ is the probability simplex, and pick $\boldsymbol{x}^0 = n^{-1}\mathbf{1}$

(2) set $\phi(\boldsymbol{x}) = +\sum_{i=1}^n x_i \log x_i$, which is 1-strongly convex w.r.t. $\|\cdot\|_1$. Then

$$\sup_{\boldsymbol{x}\in\Delta} D_\varphi(\boldsymbol{x}, \boldsymbol{x}^0) = \sup_{\boldsymbol{x}\in\Delta} \mathsf{KL}(\boldsymbol{x} \,\|\, \boldsymbol{x}^0) = \sup_{\boldsymbol{x}\in\Delta} \sum_{i=1}^n x_i \log x_i - \sum_{i=1}^n x_i \log \frac{1}{n}$$
$$= \log n + \sup_{\boldsymbol{x}\in\Delta} \sum\nolimits_{i=1}^n x_i \log x_i \leq \log n$$

Then Theorem 5.3 says

$$f^{\mathsf{best},t} - f^{\mathsf{opt}} \leq O\left(L_{f,\infty} \sqrt{\log n} \frac{\log t}{\sqrt{t}} \right)$$

if any subgradient $\boldsymbol{g}$ obeys $\|\boldsymbol{g}\|_\infty \leq L_{f,\infty}$

Example: optimization over probability simplex

Comparing these two choices and ignoring log terms, we have

$$\text{Euclidean: } O\left(\frac{L_{f,2}}{\sqrt{t}}\right) \qquad \text{vs.} \qquad \text{KL: } O\left(\frac{L_{f,\infty}}{\sqrt{t}}\right)$$

Since $\|\boldsymbol{g}\|_\infty \le \|\boldsymbol{g}\|_2 \le \sqrt{n}\|\boldsymbol{g}\|_\infty$, one has

$$\frac{1}{\sqrt{n}} \le \frac{L_{f,\infty}}{L_{f,2}} \le 1$$

and hence the KL version often yields much better performance

$$\mathrm{Loss}_{MSE} = \frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \checkmark$$

min

$$N(0,1)$$

$$p(y_i \mid x_i) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(y_i - \hat{y}_i)^2}{2}\right\}$$

$i: 1-N$

$$\max \log\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(y_i - \hat{y}_i)^2}{2}\right\}\right)$$

$$\max -\frac{N}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$\min \frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Binary: $\begin{cases} p(y_i=1|x_i) = p_i \\ p(y_i=0|x_i) = 1-p_i \end{cases}$

$\Downarrow$

$$p(y_i|x_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

$$\log \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i}$$

$\Downarrow$

$$\min -\sum_{i=1}^{N} \left(y_i \log p_i + (1-y_i)\log(1-p_i)\right)$$

$$\min -\sum_{i=1}^{N}\sum_{k=1}^{K} y_i^k \log p_i^k$$

Numerical example: robust regression

taken from Stanford EE364B

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \sum_{i=1}^{m} |\boldsymbol{a}_i^\top \boldsymbol{x} - b_i|$$

$$\text{subject to} \quad \boldsymbol{x} \in \Delta = \{\boldsymbol{x} \in \mathbb{R}_+^n \mid \boldsymbol{1}^\top \boldsymbol{x} = 1\}$$

with $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{n\times n})$ and $b_i = \frac{a_{i,1}+a_{i,2}}{2} + \mathcal{N}(0, 10^{-2})$, $m = 20$, $n = 3000$

subgradient of f: $g \in \partial f$

$$g = \sum_{i=1}^{m} \underbrace{\mathrm{sgn}(a_i^\top x - b_i)}\, a_i$$

① $\varphi(x) = \frac{1}{2}\|x\|_2^2$, $D_\varphi(x, x_t) = \frac{1}{2}\|x - x^t\|_2^2$

$$x^{t+1} = x^t - \eta_t g^t$$

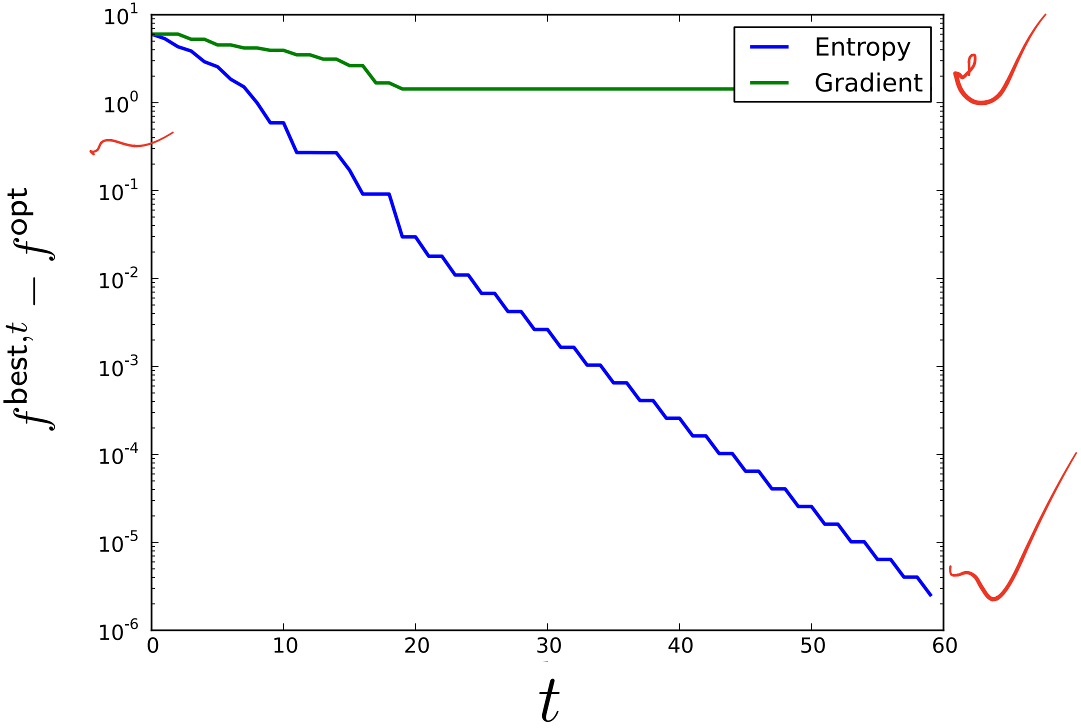
② $\varphi(x) = \sum_i x_i \log x_i$, $D_\varphi(x, x_t) = \sum_i x_i \log \frac{x_i}{x_i^t}$

$$= \sum_i x_i \log x_i - \sum_i x_i \log x_i^t$$

$$(x_i^{t+1}) = \frac{x_i^t \exp(-\eta_t g_i^t)}{\sum_j^N x_j^t \exp(-\eta_t g_j^t)}$$

Numerical example: robust regression

taken from Stanford EE364B



Fundamental inequality for mirror descent

Lemma 5.4

$$\eta_t \left(f(\boldsymbol{x}^t) - f^{\mathsf{opt}} \right) \le D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^t) - D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^{t+1}) + \frac{\eta_t^2 L_f^2}{2\rho}$$

- $D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^t) - D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^{t+1})$ motivates us to form a telescopic sum later

Proof of Lemma 5.4

$$
\begin{aligned}
f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) &\le \langle \boldsymbol{g}^t, \boldsymbol{x}^t - \boldsymbol{x}^* \rangle && \text{(property of subgradient)} \\
&= \frac{1}{\eta_t} \langle \nabla\varphi(\boldsymbol{x}^t) - \nabla\varphi(\boldsymbol{y}^{t+1}), \boldsymbol{x}^t - \boldsymbol{x}^* \rangle && \text{(MD update rule)} \\
&= \frac{1}{\eta_t} \left\{ D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^t) + D_\varphi(\boldsymbol{x}^t, \boldsymbol{y}^{t+1}) - D_\varphi(\boldsymbol{x}^*, \boldsymbol{y}^{t+1}) \right\} && \text{(three point lemma)} \\
&\le \frac{1}{\eta_t} \left\{ D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^t) + D_\varphi(\boldsymbol{x}^t, \boldsymbol{y}^{t+1}) - D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^{t+1}) - D_\varphi(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}) \right\} \\
& && \text{(Pythagorean)} \\
&= \frac{1}{\eta_t} \left\{ D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^t) - D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^{t+1}) \right\} + \frac{1}{\eta_t} \left\{ D_\varphi(\boldsymbol{x}^t, \boldsymbol{y}^{t+1}) - D_\varphi(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}) \right\}
\end{aligned}
$$

so we need to first bound the 2nd term of the last line

Proof of Lemma 5.4 (cont.)

We claim that

$$D_\varphi(\boldsymbol{x}^t, \boldsymbol{y}^{t+1}) - D_\varphi(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}) \leq \frac{(\eta_t L_f)^2}{2\rho} \tag{5.6}$$

This gives

$$\eta_t \left(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\right) \leq \left\{D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^t) - D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^{t+1})\right\} + \frac{(\eta_t L_f)^2}{2\rho}$$

as claimed

Proof of Lemma 5.4 (cont.)

Finally, we justify (5.6):

$$
\begin{aligned}
& D_\varphi\big(\boldsymbol{x}^t, \boldsymbol{y}^{t+1}\big) - D_\varphi\big(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}\big) \\
&= \varphi\big(\boldsymbol{x}^t\big) - \varphi\big(\boldsymbol{x}^{t+1}\big) - \big\langle \nabla\varphi\big(\boldsymbol{y}^{t+1}\big), \boldsymbol{x}^t - \boldsymbol{x}^{t+1} \big\rangle \checkmark \\
&\leq \big\langle \nabla\varphi\big(\boldsymbol{x}^t\big), \boldsymbol{x}^t - \boldsymbol{x}^{t+1} \big\rangle - \frac{\rho}{2}\big\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\big\|^2 - \big\langle \nabla\varphi\big(\boldsymbol{y}^{t+1}\big), \boldsymbol{x}^t - \boldsymbol{x}^{t+1} \big\rangle \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(strong convexity of } \varphi\text{)} \\
&= \big\langle \nabla\varphi\big(\boldsymbol{x}^t\big) - \nabla\varphi\big(\boldsymbol{y}^{t+1}\big), \boldsymbol{x}^t - \boldsymbol{x}^{t+1} \big\rangle - \frac{\rho}{2}\big\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\big\|_2^2 \\
&= \eta_t \langle \boldsymbol{g}^t, \boldsymbol{x}^t - \boldsymbol{x}^{t+1} \rangle - \frac{\rho}{2}\big\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\big\|^2 \qquad \text{(MD update rule)} \\
&\leq \eta_t L_f \big\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\big\| - \frac{\rho}{2}\big\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\big\|^2 \qquad \text{(Cauchy-Schwarz)} \\
&\leq \frac{(\eta_t L_f)^2}{2\rho} \qquad \text{(optimize quadratic function in } \|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|\text{)}
\end{aligned}
$$

$$\|x^t - x^{t+1}\|^2 = \frac{\eta_t L_f}{\rho}$$

Proof of Theorem 5.3

From Lemma 5.4, one has

$$\eta_k \left(f(\boldsymbol{x}^k) - f^{\mathsf{opt}} \right) \leq D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^k) - D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^{k+1}) + \frac{\eta_k^2 L_f^2}{2\rho}$$

Taking this inequality for $k = 0, \cdots, t$ and summing them up give

$$\begin{aligned}
\sum_{k=0}^{t} \eta_k \left(f(\boldsymbol{x}^k) - f^{\mathsf{opt}} \right) &\leq D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^0) - D_\varphi(\boldsymbol{x}^*, \boldsymbol{x}^{t+1}) + \frac{L_f^2 \sum_{k=0}^{t} \eta_k^2}{2\rho} \\
&\leq \sup_{\boldsymbol{x} \in \mathcal{C}} D_\varphi(\boldsymbol{x}, \boldsymbol{x}^0) + \frac{L_f^2 \sum_{k=0}^{t} \eta_k^2}{2\rho}
\end{aligned}$$

This together with $f^{\mathsf{best},t} - f^{\mathsf{opt}} \leq \frac{\sum_{k=0}^{t} \eta_k \left(f(\boldsymbol{x}^k) - f^{\mathrm{opt}} \right)}{\sum_{k=0}^{t} \eta_k}$ concludes the proof

Reference

[1] "*Problem complexity and method efficiency in optimization*," A. Nemirovski, D. Yudin, Wiley, 1983.

[2] "*Mirror descent and nonlinear projected subgradient methods for convex optimization*," A. Beck, M. Teboulle, Operations Research Letters, 31(3), 2003.

[3] "*Convex optimization: algorithms and complexity*," S. Bubeck, Foundations and trends in machine learning, 2015.

[4] "*First-order methods in optimization*," A. Beck, Vol. 25, SIAM, 2017.

[5] "*Mathematical optimization, MATH301 lecture notes*," E. Candes, Stanford.

[6] "*Convex optimization, EE364B lecture notes*," S. Boyd, Stanford.

Reference

[7] "*Matrix nearness problems with Bregman divergences*," I. Dhillon, J. Tropp, SIAM Journal on Matrix Analysis and Applications, 29(4), 2007.

[8] "*Nonlinear Programming (2nd Edition)*," D. Bertsekas, Athena Scientific, 2016.