# Subgradient methods

**Materials from**

- A. Beck, M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. ORL03.
- Y. Nesterov. Primal-dual subgradient methods for convex problems. MP09.
- Y. Nesterov, V. Shikhman. Quasi-monotone subgradient methods for nonsmooth convex minimization. JOTA15.

# Projected subgradient method

Consider the following nonsmooth convex minimization problem,

$$(P) \quad \text{minimize } f(x) \text{ s.t. } x \in X \subset \mathbb{R}^n.$$

**Subgradient:** A subgradient of $f$ at $x \in X$ is computable. An element of the subdifferential $\partial f(x)$ is denoted by $f'(x)$.

**Update formula:**

$$x_{k+1} = \pi_X(x^k - t_k f'(x^k)),$$

$$t_k > 0 \text{ (a stepsize)}, \tag{1.1}$$

where $\pi_X(x) = \operatorname{argmin}\{\|x - y\| \mid y \in X\}$ is the Euclidean projection onto $X$.

# Mirror descent algorithm

Let $\psi : X \to \mathbb{R}$ be strongly convex and continuously differentiable on $\operatorname{int} X$. The distance-like function is defined by $B_\psi : X \times \operatorname{int}(X) \to \mathbb{R}$ given by

$$B_\psi(x, y) = \psi(x) - \psi(y) - \langle x - y, \nabla\psi(y)\rangle. \qquad (3.10)$$

$$x^{k+1} = \operatorname*{argmin}_{x \in X} \left\{ \langle x, f'(x^k)\rangle + \frac{1}{t_k} B_\psi(x, x^k) \right\},$$

$$t_k > 0. \qquad (3.11)$$

arbitrary point $x^1 \in \mathbb{R}^n$. With $X = \mathfrak{R}^n$ and $\psi(x) = \frac{1}{2}\|x\|^2$ one obtains $B_\psi(x, y) = \frac{1}{2}\|x - y\|^2$ thus recovering the classical squared Euclidean distance and SANP is just the classical subgradient algorithm.

# Convergence result

$$\sum_{k=1}^{s} t_k(f(x^k) - f(x^*)) \leqslant B_\psi(x^*, x^1) - B_\psi(x^*, x^{s+1}) + (2\sigma)^{-1} \sum_{k=1}^{s} t_k^2 \|f'(x^k)\|_*^2.$$

$$t_k := \frac{\sqrt{2\sigma B_\psi(x^*, x^1)}}{L_f} \frac{1}{\sqrt{k}}, \tag{4.23}$$

*one has the following efficiency estimate*

$$\min_{1 \leqslant s \leqslant k} f(x^s) - \min_{x \in X} f(x) \leqslant L_f \sqrt{\frac{2B_\psi(x^*, x^1)}{\sigma}} \frac{1}{\sqrt{k}}. \tag{4.24}$$

# A simple proof for Euclidean setting

$$\|x - x_{k+1}\|_2^2 \overset{(1.3)}{=} \|x - x_k\|_2^2 + 2\lambda_k \langle g_k, x - x_k \rangle + \lambda_k^2 \|g_k\|_2^2$$
$$\leq \|x - x_k\|_2^2 + 2\lambda_k \langle g_k, x - x_k \rangle + \lambda_k^2 L^2.$$

for any $x \in R^n$ with $\frac{1}{2}\|x - x_0\|_2^2 \leq D$

$$f(x) \geq l_k(x) \overset{\text{def}}{=} \sum_{i=0}^{k} \lambda_i [f(x_i) + \langle g_i, x - x_i \rangle] / \sum_{i=0}^{k} \lambda_i$$
$$\geq \left\{ \sum_{i=0}^{k} \lambda_i f(x_i) - D - \frac{1}{2} L^2 \sum_{i=0}^{k} \lambda_i^2 \right\} / \sum_{i=0}^{k} \lambda_i .$$

**From:** Y. Nesterov. Primal-dual subgradient methods for convex problems. MP09.

# A simple proof for Euclidean setting (cont.)

Take $$\lambda_k > 0, \quad \lambda_k \to 0, \quad \sum_{k=0}^{\infty} \lambda_k = \infty. \tag{1.2}$$

Thus, denoting $f_D^* = \min_x \{ f(x) : \ \frac{1}{2}\|x - x_0\|_2^2 \leq D \}$,

$$\bar{f}_k = \frac{\sum_{i=0}^{k} \lambda_i f(x_i)}{\sum_{i=0}^{k} \lambda_i}, \quad \omega_k = \frac{2D + L^2 \sum_{i=0}^{k} \lambda_i^2}{2 \sum_{i=0}^{k} \lambda_i},$$

we conclude that $\bar{f}_k - f_D^* \leq \omega_k$. Note that the conditions (1.2) are necessary and sufficient for $\omega_k \to 0$.

# Dual averaging scheme

**Motivation: it is noticed that for previous method,**

*New subgradients enter the model with decreasing weights.*

| |
|---|
| **Initialization:** Set $s_0 = 0 \in E^*$. Choose $\beta_0 > 0$. |
| **Iteration** ($k \geq 0$):<br>**1.** Compute $g_k = \mathcal{G}(x_k)$.<br>**2.** Choose $\lambda_k > 0$. Set $s_{k+1} = s_k + \lambda_k g_k$.<br>**3.** Choose $\beta_{k+1} \geq \beta_k$. Set $x_{k+1} = \pi_{\beta_{k+1}}(-s_{k+1})$. |

$$\min_x \{ f(x) : \ x \in Q \},$$

$Q$ be a closed convex set in $E$

a black-box oracle $\mathcal{G}(\cdot)$

$$s_{k+1} = \sum_{i=0}^{k} \lambda_i g_i ,$$

a prox-function $d(x)$

$$\pi_\beta(s) \stackrel{\text{def}}{=} \arg\min_{x \in Q} \{ -\langle s, x \rangle + \beta d(x) \}$$

# Convergence result of dual averaging

**Theorem 1** *Let the sequences $X_k$, $G_k$ and $\Lambda_k$ be generated by (2.14). Then:*

1. *For any $k \geq 0$ and $D \geq 0$ we have:*

$$\delta_k(D) \leq \Delta_k(\beta_{k+1}, D) \leq \beta_{k+1} D + \frac{1}{2\sigma} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_i} \|g_i\|_*^2. \tag{2.15}$$

2. *Assume that a solution $x^*$ in the sense (2.8) exists. Then*

$$\tfrac{1}{2}\sigma \|x_{k+1} - x^*\|^2 \leq d(x^*) + \frac{1}{2\sigma\beta_{k+1}} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_i} \|g_i\|_*^2. \tag{2.16}$$

where $\quad \delta_k(D) = \max_x \left\{ \sum_{i=0}^{k} \lambda_i \langle g_i, x_i - x \rangle : \; x \in \mathcal{F}_D, \right\}, \quad D \geq 0.$

$$\mathcal{F}_D = \{x \in Q : \; d(x) \leq D\}$$

# Proof of Theorem 1

Let
$$\begin{aligned} \xi_D(s) &= \max_{x\in Q}\{\langle s, x - x_0\rangle :\ d(x) \le D\}, \\ V_\beta(s) &= \max_{x\in Q}\{\langle s, x - x_0\rangle - \beta d(x)\}, \end{aligned} \tag{2.1}$$

Then
$$V_{\beta_2}(s) \le V_{\beta_1}(s). \tag{2.2}$$

$$\nabla V_\beta(s) = \pi_\beta(s) - x_0, \quad \pi_\beta(s) \stackrel{\text{def}}{=} \arg\min_{x\in Q}\{-\langle s, x\rangle + \beta d(x)\}. \tag{2.4}$$

$$V_\beta(s+\delta) \le V_\beta(s) + \langle \delta, \nabla V_\beta(s)\rangle + \frac{1}{2\sigma\beta}\|\delta\|_*^2 \quad \forall s, \delta \in E^*. \tag{2.5}$$

and
$$\delta_k(D) = \sum_{i=0}^{k} \lambda_i \langle g_i, x_i - x_0\rangle + \xi_D(-s_{k+1}). \tag{2.11}$$

# Proof of Theorem 1 (cont.)

Note
$$
\begin{aligned}
V_{\beta_{i+1}}(-s_{i+1}) &\overset{(2.2)}{\leq} V_{\beta_i}(-s_{i+1}) \\
&\overset{(2.5)}{\leq} V_{\beta_i}(-s_i) - \lambda_i \langle g_i, \nabla V_{\beta_i}(-s_i)\rangle + \frac{\lambda_i^2}{2\sigma\beta_i}\|g_i\|_*^2 \\
&\overset{(2.4)}{=} V_{\beta_i}(-s_i) + \lambda_i \langle g_i, x_0 - x_i\rangle + \frac{\lambda_i^2}{2\sigma\beta_i}\|g_i\|_*^2.
\end{aligned}
$$

Thus,

$$\lambda_i \langle g_i, x_i - x_0\rangle \leq V_{\beta_i}(-s_i) - V_{\beta_{i+1}}(-s_{i+1}) + \frac{\lambda_i^2}{2\sigma\beta_i}\|g_i\|_*^2, \quad i = 1, \ldots, k.$$

The summation of all these inequalities results in

$$\sum_{i=0}^{k} \lambda_i \langle g_i, x_i - x_0\rangle \leq V_{\beta_1}(-s_1) - V_{\beta_{k+1}}(-s_{k+1}) + \frac{1}{2\sigma}\sum_{i=1}^{k}\frac{\lambda_i^2}{\beta_i}\|g_i\|_*^2. \quad (2.18)$$

But in view of (2.6) $V_{\beta_1}(-s_1) \leq \frac{\lambda_0^2}{2\sigma\beta_1}\|g_0\|_*^2 \leq \frac{\lambda_0^2}{2\sigma\beta_0}\|g_0\|_*^2$. Thus, (2.18) results in (2.15).

# Simple dual averaging

**Initialization:** Set $s_0 = 0 \in E^*$. Choose $\gamma > 0$.

**Iteration** ($k \geq 0$):

**1.** Compute $g_k = \mathcal{G}(x_k)$. Set $s_{k+1} = s_k + g_k$.

**2.** Choose $\beta_{k+1} = \gamma \hat{\beta}_{k+1}$. Set $x_{k+1} = \pi_{\beta_{k+1}}(-s_{k+1})$.

*Simple averages*: $\lambda_k = 1$

and take $\hat{\beta}_0 = \hat{\beta}_1 = 1, \quad \hat{\beta}_{k+1} = \sum_{i=0}^{k} \frac{1}{\hat{\beta}_i}, \quad k \geq 0.$

**Lemma 3:** $\sqrt{2k-1} \leq \hat{\beta}_k \leq \frac{1}{1+\sqrt{3}} + \sqrt{2k-1}, \quad k \geq 1.$

# Convergence of simple dual averaging

**Theorem 2** *Assume that* $\|g_k\|_* \le L$, $k \ge 0$. *For method* (2.21) *we have* $S_k = k + 1$ *and*

$$\delta_k(D) \le \hat{\beta}_{k+1}\left(\gamma D + \frac{L^2}{2\sigma\gamma}\right).$$

Note $$\frac{1}{S_k}\delta_k(D) = \frac{1}{S_k}\sum_{i=0}^{k}\lambda_i f(x_i) - \hat{f}_N(D) \ge f(\hat{x}_{k+1}) - f_D^*. \qquad (3.2)$$

**Simple averages.** In view of Theorem 2 and inequalities (2.20), (3.2) we have

$$f(\hat{x}_{k+1}) - f_D^* \le \frac{0.5 + \sqrt{2k+1}}{k+1}\left(\gamma D + \frac{L^2}{2\sigma\gamma}\right). \qquad (3.3)$$

# Weighted dual averaging

**Initialization:** Set $s_0 = 0 \in E^*$. Choose $\rho > 0$.

**Iteration** ($k \geq 0$)**:**

**1.** Compute $g_k = \mathcal{G}(x_k)$. Set $s_{k+1} = s_k + g_k/\|g_k\|_*$.

**2.** Choose $\beta_{k+1} = \frac{\hat{\beta}_{k+1}}{\rho\sqrt{\sigma}}$. Set $x_{k+1} = \pi_{\beta_{k+1}}(-s_{k+1})$.

*Weighted averages*: $\lambda_k = \frac{1}{\|g_k\|_*}$

and take $\hat{\beta}_0 = \hat{\beta}_1 = 1, \ \hat{\beta}_{k+1} = \sum_{i=0}^{k} \frac{1}{\hat{\beta}_i}, \quad k \geq 0.$

# Convergence of weighted dual averaging

**Theorem 3** *Assume that* $\|g_k\|_* \leq L$, $k \geq 0$. *For method* (2.22) *we have* $S_k \geq \frac{k+1}{L}$ *and*

$$\delta_k(D) \leq \frac{\hat{\beta}_{k+1}}{\sqrt{\sigma}} \left( \frac{D}{\rho} + \tfrac{1}{2}\rho \right).$$

Again note $$\frac{1}{S_k}\delta_k(D) = \frac{1}{S_k}\sum_{i=0}^{k} \lambda_i f(x_i) - \hat{f}_N(D) \geq f(\hat{x}_{k+1}) - f_D^*. \tag{3.2}$$

**Weighted averages.** In view of Theorem 3 and inequalities (2.20), (3.2) we have

$$f(\hat{x}_{k+1}) - f_D^* \leq \frac{0.5 + \sqrt{2k+1}}{(k+1)\sqrt{\sigma}} L \left( \frac{1}{\rho} D + \frac{\rho}{2} \right). \tag{3.5}$$

# Double averaging

**Motivation:**

Recently, it became clear that all methods mentioned above have a common drawback:

*They cannot generate a convergent sequence of test points.*

| **Subgradient Method with Double Averaging** |
| --- |
| **1**. Compute $x_t^+ = \arg\min_{x \in Q} \{A_t \langle s_t, x \rangle + \gamma_t d(x)\}$. <br> **2**. Define $\tau_t = \frac{a_{t+1}}{A_{t+1}}$. Update $x_{t+1} = (1 - \tau_t) x_t + \tau_t x_t^+$. |

where $A_t = \sum_{k=0}^{t} a_k$

$$s_t = \frac{1}{A_t} \sum_{k=0}^{t} a_k \nabla f(x_k)$$

# Convergence analysis of double averaging

**Goal:** $$A_t f(x_t) \leq \sum_{k=0}^{t} a_k[f(x_k) + \langle \nabla f(x_k), x - x_k \rangle] + \gamma_t d(x) + B_t \quad \forall x \in Q, \qquad (22)$$

**Corollary 3.1** *Let a sequence of points* $\{x_t\}_{t\geq 0}$ *satisfy condition (22). Then for any* $t \geq 0$ *we have*

$$f(x_t) - f_* + \|s_t\|_R^* \leq \frac{1}{A_t}(B_t + \gamma_t G_R), \qquad (27)$$

*where* $G_R = \max_{x\in Q}\{d(x) : \ \|x - x^*\| \leq R\}$.

$$\|s\|_R^* = \max_{x\in Q}\{\langle s, x_* - x\rangle : \ \|x - x_*\| \leq R\}, \quad s \in \mathbb{E}^*.$$

# Proof of Corollary 3.1

*Proof* In view of condition (22), for any $x \in Q$ and $y \in \mathbb{E}$, we have

$$\begin{aligned}
&\sum_{k=0}^{t} a_k f(x_k) + \gamma_t d(x) + B_t \\
&\quad \geq A_t f(x_t) + A_t \langle s_t, y - x \rangle + \sum_{k=0}^{t} a_k \langle \nabla f(x_k), x_k - y \rangle \\
&\quad \overset{(18)}{\geq} A_t f(x_t) + A_t \langle s_t, y - x \rangle + \sum_{k=0}^{t} a_k f(x_k) - A_t f(y).
\end{aligned}$$

Thus, $\frac{1}{A_t}(B_t + \gamma_t d(x)) \geq f(x_t) + [\langle s_t, y \rangle - f(y)] + \langle -s_t, x \rangle$

Let us choose $C = \{x \in Q : \ \|x - x_*\| \leq R\}$

Then $$\frac{1}{A_t}(B_t + \gamma_t G_R) \geq f(x_t) + \langle s_t, x_* \rangle - f_* + \langle -s_t, x \rangle, \quad x \in C.$$

Maximizing the right-hand side of this inequality in $x$, we obtain (27) from (26).

# Convergence analysis of double averaging

**Goal:** $$A_t f(x_t) \le \sum_{k=0}^{t} a_k[f(x_k) + \langle \nabla f(x_k), x - x_k \rangle] + \gamma_t d(x) + B_t \quad \forall x \in Q, \qquad (22)$$

**Theorem 3.1** *Let the sequence $\{x_t\}_{t\ge 0}$ be generated by method (28) with monotone sequence of parameters $\{\gamma_t\}_{t\ge 0}$:*

$$\gamma_{t+1} \ge \gamma_t, \quad t \ge 0. \qquad (30)$$

*Then condition (22) holds with*

$$B_t = \frac{1}{2} \sum_{k=0}^{t} \frac{a_k^2}{\gamma_{k-1}} \|\nabla f(x_k)\|_*^2, \qquad (31)$$

*where $\gamma_{-1} = \gamma_0$.*

**Proof by induction**

# Double simple averaging

| **Subgradient Method with Double Simple Averaging** |
| --- |
| **1**. Compute $x_t^+ = \arg\min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} \nabla f(x_k), x \rangle + \gamma_t d(x) \right\}$. <br> **2**. Update $x_{t+1} = \frac{t+1}{t+2} x_t + \frac{1}{t+2} x_t^+$. |

**Set** $a_t = 1, t \geq 0$ **In double averaging scheme**

# Convergence of simple double averaging

**Theorem 3.2** *Let sequence* $\{x_t\}_{t\geq 0}$ *be generated by method (34) with parameters* $\{\gamma_t\}_{t\geq 0}$ *satisfying condition (30). Then, for any* $t \geq 0$*, we have*

$$f(x_t) - f_* + \|s_t\|_R^* \leq \frac{1}{t+1}\left(\gamma_t G_R + \tfrac{1}{2}\sum_{k=0}^{t} \frac{\|\nabla f(x_k\|_*^2}{\gamma_{k-1}}\right). \tag{35}$$

$$\|\nabla f(x)\|_* \leq L, \quad x \in \operatorname{int} Q. \tag{36}$$

**Corollary 3.3** *Assume that in method (34) we have*

$$\gamma_t \to \infty, \quad \frac{\gamma_t}{t+1} \to 0. \tag{37}$$

*Then* $\lim\limits_{t\to\infty} f(x_t) = f_*$ *and* $\lim\limits_{t\to\infty} \|s_t\|_R^* = 0.$

# Convergence rate of simple double averaging

Take $$\gamma_t = \gamma\sqrt{t+1}, \quad t \geq 0, \tag{38}$$

**Corollary 3.4** *Let objective function of problem (17) satisfy condition (36), and the sequence* $\{\gamma_t\}_{t\geq 0}$ *be defined by the rule (38). Then, for any* $t \geq 0$, *we have*

$$f(x_t) - f_* + \|s_t\|_R^* \leq \frac{1}{\sqrt{t+1}}\left(\gamma G_R + \frac{1}{\gamma}L^2\right),$$