Stochastic Subgradient Method

- noisy unbiased subgradient
- stochastic subgradient method
- convergence proof
- stochastic programming
- expected value of convex function
- on-line learning and adaptive signal processing

Noisy unbiased subgradient

- random vector $\tilde{g} \in \mathbf{R}^n$ is a **noisy unbiased subgradient** for $f : \mathbf{R}^n \to \mathbf{R}$ at x if for all z

$$f(z) \geq f(x) + (\mathbf{E}\,\tilde{g})^T(z - x)$$

  i.e., $g = \mathbf{E}\,\tilde{g} \in \partial f(x)$

- same as $\tilde{g} = g + v$, where $g \in \partial f(x)$, $\mathbf{E}\,v = 0$

- v can represent error in computing g, measurement noise, Monte Carlo sampling error, etc.

- if x is also random, $\tilde{g}$ is a noisy unbiased subgradient of f at x if

$$\forall z \qquad f(z) \geq f(x) + \mathbf{E}(\tilde{g}|x)^T(z - x)$$

  holds almost surely

- same as $\mathbf{E}(\tilde{g}|x) \in \partial f(x)$ (a.s.)

Stochastic subgradient method

stochastic subgradient method is the subgradient method, using noisy unbiased subgradients

$$x^{(k+1)} = x^{(k)} - \alpha_k \tilde{g}^{(k)}$$

- $x^{(k)}$ is kth iterate
- $\tilde{g}^{(k)}$ is any noisy unbiased subgradient of (convex) f at $x^{(k)}$, *i.e.*,

$$\mathbf{E}(\tilde{g}^{(k)}|x^{(k)}) = g^{(k)} \in \partial f(x^{(k)})$$

- $\alpha_k > 0$ is the kth step size
- define $f_{\rm best}^{(k)} = \min\{f(x^{(1)}), \ldots, f(x^{(k)})\}$

Assumptions

- $f^\star = \inf_x f(x) > -\infty$, with $f(x^\star) = f^\star$
- $\mathbf{E}\,\|g^{(k)}\|_2^2 \leq G^2$ for all k
- $\mathbf{E}\,\|x^{(1)} - x^\star\|_2^2 \leq R^2$ (can take $=$ here)
- step sizes are square-summable but not summable

$$\alpha_k \geq 0, \qquad \sum_{k=1}^{\infty} \alpha_k^2 = \|\alpha\|_2^2 < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

these assumptions are stronger than needed, just to simplify proofs

Convergence results

- convergence in expectation:

$$\lim_{k\to\infty} \mathbf{E}\, f_{\rm best}^{(k)} = f^\star$$

- convergence in probability: for any $\epsilon > 0$,

$$\lim_{k\to\infty} \mathbf{Prob}(f_{\rm best}^{(k)} \geq f^\star + \epsilon) = 0$$

- almost sure convergence:

$$\lim_{k\to\infty} f_{\rm best}^{(k)} = f^\star$$

  a.s. (we won't show this)

Convergence proof

key quantity: *expected Euclidean distance squared to the optimal set*

$$\mathbf{E}\left(\|x^{(k+1)} - x^\star\|_2^2 \;\middle|\; x^{(k)}\right) = \mathbf{E}\left(\|x^{(k)} - \alpha_k \tilde{g}^{(k)} - x^\star\|_2^2 \;\middle|\; x^{(k)}\right)$$

$$\begin{aligned}
&= \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k\,\mathbf{E}\left(\tilde{g}^{(k)T}(x^{(k)} - x^\star) \;\middle|\; x^{(k)}\right) + \alpha_k^2\,\mathbf{E}\left(\|\tilde{g}^{(k)}\|_2^2 \;\middle|\; x^{(k)}\right) \\
&= \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k\,\mathbf{E}(\tilde{g}^{(k)}|x^{(k)})^T(x^{(k)} - x^\star) + \alpha_k^2\,\mathbf{E}\left(\|\tilde{g}^{(k)}\|_2^2 \;\middle|\; x^{(k)}\right) \\
&\leq \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k(f(x^{(k)}) - f^\star) + \alpha_k^2\,\mathbf{E}\left(\|\tilde{g}^{(k)}\|_2^2 \;\middle|\; x^{(k)}\right)
\end{aligned}$$

using $\mathbf{E}(\tilde{g}^{(k)}|x^{(k)}) \in \partial f(x^{(k)})$

now take expectation:

$$\mathbf{E}\,\|x^{(k+1)} - x^\star\|_2^2 \le \mathbf{E}\,\|x^{(k)} - x^\star\|_2^2 - 2\alpha_k(\mathbf{E}\,f(x^{(k)}) - f^\star) + \alpha_k^2\,\mathbf{E}\,\|\tilde{g}^{(k)}\|_2^2$$

apply recursively, and use $\mathbf{E}\,\|\tilde{g}^{(k)}\|_2^2 \le G^2$ to get

$$\mathbf{E}\,\|x^{(k+1)} - x^\star\|_2^2 \le \mathbf{E}\,\|x^{(1)} - x^\star\|_2^2 - 2\sum_{i=1}^k \alpha_i(\mathbf{E}\,f(x^{(i)}) - f^\star) + G^2\sum_{i=1}^k \alpha_i^2$$

and so

$$\min_{i=1,\ldots,k}(\mathbf{E}\,f(x^{(i)}) - f^\star) \le \frac{R^2 + G^2\|\alpha\|_2^2}{2\sum_{i=1}^k \alpha_i}$$

- we conclude $\min_{i=1,\ldots,k} \mathbf{E}\, f(x^{(i)}) \to f^\star$

- Jensen's inequality and concavity of minimum yields

$$\mathbf{E}\, f_{\rm best}^{(k)} = \mathbf{E} \min_{i=1,\ldots,k} f(x^{(i)}) \leq \min_{i=1,\ldots,k} \mathbf{E}\, f(x^{(i)})$$

so $\mathbf{E}\, f_{\rm best}^{(k)} \to f^\star$ (convergence in expectation)

- Markov's inequality: for $\epsilon > 0$

$$\mathbf{Prob}(f_{\rm best}^{(k)} - f^\star \geq \epsilon) \leq \frac{\mathbf{E}(f_{\rm best}^{(k)} - f^\star)}{\epsilon}$$

righthand side goes to zero, so we get convergence in probability

Example

piecewise linear minimization

$$\text{minimize} \quad f(x) = \max_{i=1,\ldots,m}(a_i^T x + b_i)$$

we use stochastic subgradient algorithm with noisy subgradient

$$\tilde{g}^{(k)} = g^{(k)} + v^{(k)}, \qquad g^{(k)} \in \partial f(x^{(k)})$$

$v^{(k)}$ independent zero mean random variables

problem instance: $n = 20$ variables, $m = 100$ terms, $f^\star \approx 1.1$, $\alpha_k = 1/k$
$v^{(k)}$ are IID $\mathcal{N}(0, 0.5I)$ (25% noise since $\|g\| \approx 4.5$)