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 \( \tilde{g} \in \mathbb{R}^n \) is a noisy unbiased subgradient for \( f : \mathbb{R}^n \to \mathbb{R} \) at \( x \) if for all \( z \)

\[
f(z) \geq f(x) + (E\tilde{g})^T(z - x)
\]

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

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

• \( v \) 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

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

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

\[ \mathbb{E}(\tilde{g}^{(k)}|x^{(k)}) = g^{(k)} \in \partial f(x^{(k)}) \]

- \( \alpha_k > 0 \) is the \( k \)th step size
- define \( f^{(k)}_{\text{best}} = \min\{f(x^{(1)}), \ldots, f(x^{(k)})\} \)
Assumptions

• $f^* = \inf_x f(x) > -\infty$, with $f(x^*) = f^*$

• $\mathbb{E} \|g^{(k)}\|_2^2 \leq G^2$ for all $k$

• $\mathbb{E} \|x^{(1)} - x^*\|_2^2 \leq R^2$ (can take $= \text{here}$)

• step sizes are square-summable but not summable

\[
\alpha_k \geq 0, \quad \sum_{k=1}^{\infty} \alpha_k^2 = \|\alpha\|_2^2 < \infty, \quad \sum_{k=1}^{\infty} \alpha_k = \infty
\]

these assumptions are stronger than needed, just to simplify proofs
Convergence results

• convergence in expectation:

\[
\lim_{k \to \infty} \mathbf{E} f_{\text{best}}^{(k)} = f^*
\]

• convergence in probability: for any \( \epsilon > 0 \),

\[
\lim_{k \to \infty} \mathbf{Prob}(f_{\text{best}}^{(k)} \geq f^* + \epsilon) = 0
\]

• almost sure convergence:

\[
\lim_{k \to \infty} f_{\text{best}}^{(k)} = f^*
\]

a.s. (we won’t show this)
Convergence proof

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

\[
E \left( \| x^{(k+1)} - x^* \|_2^2 | x^{(k)} \right) = E \left( \| x^{(k)} - \alpha_k \tilde{g}^{(k)} - x^* \|_2^2 | x^{(k)} \right)
\]

\[
= \| x^{(k)} - x^* \|_2^2 - 2\alpha_k E \left( (\tilde{g}^{(k)})^T (x^{(k)} - x^*) | x^{(k)} \right) + \alpha_k^2 E \left( \| \tilde{g}^{(k)} \|_2^2 | x^{(k)} \right)
\]

\[
= \| x^{(k)} - x^* \|_2^2 - 2\alpha_k E(\tilde{g}^{(k)} | x^{(k)})^T (x^{(k)} - x^*) + \alpha_k^2 E \left( \| \tilde{g}^{(k)} \|_2^2 | x^{(k)} \right)
\]

\[
\leq \| x^{(k)} - x^* \|_2^2 - 2\alpha_k (f(x^{(k)}) - f^*) + \alpha_k^2 E \left( \| \tilde{g}^{(k)} \|_2^2 | x^{(k)} \right)
\]

using \( E(\tilde{g}^{(k)} | x^{(k)}) \in \partial f(x^{(k)}) \)
now take expectation:

\[ E \| x^{(k+1)} - x^* \|_2^2 \leq E \| x^{(k)} - x^* \|_2^2 - 2\alpha_k (E f(x^{(k)}) - f^*) + \alpha_k^2 E \| \tilde{g}^{(k)} \|_2^2 \]

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

\[ E \| x^{(k+1)} - x^* \|_2^2 \leq E \| x^{(1)} - x^* \|_2^2 - 2 \sum_{i=1}^{k} \alpha_i (E f(x^{(i)}) - f^*) + G^2 \sum_{i=1}^{k} \alpha_i^2 \]

and so

\[ \min_{i=1,\ldots,k} (E f(x^{(i)}) - f^*) \leq \frac{R^2 + G^2 \| \alpha \|_2^2}{2 \sum_{i=1}^{k} \alpha_i} \]
• we conclude \( \min_{i=1,\ldots,k} \mathbb{E} f(x^{(i)}) \to f^* \)

• Jensen’s inequality and concavity of minimum yields

\[
\mathbb{E} f_{\text{best}}^{(k)} = \mathbb{E} \min_{i=1,\ldots,k} f(x^{(i)}) \leq \min_{i=1,\ldots,k} \mathbb{E} f(x^{(i)})
\]

so \( \mathbb{E} f_{\text{best}}^{(k)} \to f^* \) (convergence in expectation)

• Markov’s inequality: for \( \epsilon > 0 \)

\[
\text{Prob}(f_{\text{best}}^{(k)} - f^* \geq \epsilon) \leq \frac{\mathbb{E}(f_{\text{best}}^{(k)} - f^*)}{\epsilon}
\]

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

piecewise linear minimization

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

we use stochastic subgradient algorithm with noisy subgradient

\[
\tilde{g}^{(k)} = g^{(k)} + v^{(k)}, \quad g^{(k)} \in \partial f(x^{(k)})
\]

\(v^{(k)}\) independent zero mean random variables
problem instance: $n = 20$ variables, $m = 100$ terms, $f^* \approx 1.1$, $\alpha_k = 1/k$ $v^{(k)}$ are IID $\mathcal{N}(0, 0.5I)$ (25% noise since $\|g\| \approx 4.5$)
average and one std. dev. for $f^{(k)}_{\text{best}} - f^*$ over 100 realizations
empirical distributions of $f_{\text{best}}^{(k)} - f^*$ at $k = 250$, $k = 1000$, and $k = 5000$
Stochastic programming

\[
\text{minimize } \mathbb{E} f_0(x, \omega) \\
\text{subject to } \mathbb{E} f_i(x, \omega) \leq 0, \quad i = 1, \ldots, m
\]

if \( f_i(x, \omega) \) is convex in \( x \) for each \( \omega \), problem is convex

‘certainty-equivalent’ problem

\[
\text{minimize } f_0(x, \mathbb{E} \omega) \\
\text{subject to } f_i(x, \mathbb{E} \omega) \leq 0, \quad i = 1, \ldots, m
\]

(if \( f_i(x, \omega) \) is convex in \( \omega \), gives a lower bound on optimal value of stochastic problem)
Variations

• in place of $\mathbb{E} f_i(x, \omega) \leq 0$ (constraint holds in expectation) can use

  - $\mathbb{E} f_i(x, \omega)_+ \leq \epsilon$ (LHS is expected violation)
  - $\mathbb{E} \left( \max_i f_i(x, \omega)_+ \right) \leq \epsilon$ (LHS is expected worst violation)

• unfortunately, *chance constraint* $\text{Prob}(f_i(x, \omega) \leq 0) \geq \eta$ is convex only in a few special cases
**Expected value of convex function**

suppose $F(x, w)$ is convex in $x$ for each $w$ and $G(x, w) \in \partial_x F(x, w)$

- $f(x) = \mathbb{E} F(x, w) = \int F(x, w)p(w) \, dw$ is convex
- a subgradient of $f$ at $x$ is
  \[
g = \mathbb{E} G(x, w) = \int G(x, w)p(w) \, dw \in \partial f(x)
  \]
- a noisy unbiased subgradient of $f$ at $x$ is
  \[
  \tilde{g} = \frac{1}{M} \sum_{i=1}^{M} G(x, w_i)
  \]
  where $w_1, \ldots, w_M$ are $M$ independent samples (Monte Carlo)
**Example: Expected value of piecewise linear function**

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

where \( a_i \) and \( b_i \) are random

evaluate noisy subgradient using Monte Carlo method with \( M \) samples, and run stochastic subgradient method

compare to:

- certainty equivalent: minimize \( f_{ce}(x) = \max_{i=1, \ldots, m}(\mathbb{E} a_i^T x + \mathbb{E} b_i) \)
- heuristic: minimize \( f_{heur}(x) = \max_{i=1, \ldots, m}(\mathbb{E} a_i^T x + \mathbb{E} b_i + \lambda \|x\|_2) \)
problem instance: $n = 20$, $m = 100$, $a_i \sim \mathcal{N}(\bar{a}_i, 5I)$, $b \sim \mathcal{N}(\bar{b}, 5I)$, $\|a_i\|_2 \approx 5$, $\|b\|_2 \approx 10$, $x_{\text{stoch}}$ computed using $M = 100$
$f^* \approx 1.34$ estimated by running the method with $M = 1000$ for long time

![Graph showing $f(k) - \hat{f}^*$ vs. $k$ for different values of $M$.]
On-line learning and adaptive signal processing

- \((x, y) \in \mathbb{R}^n \times \mathbb{R}\) have some joint distribution

- find weight vector \(w \in \mathbb{R}^n\) for which \(w^T x\) is a good estimator of \(y\)

- choose \(w\) to minimize expected value of a convex loss function \(l\)

\[
J(w) = \mathbb{E} l(w^T x - y)
\]

- \(l(u) = u^2\): mean-square error
- \(l(u) = |u|\): mean-absolute error

- at each step (e.g., time sample), we are given a sample \((x^{(k)}, y^{(k)})\) from the distribution
noisy unbiased subgradient of $J$ at $w^{(k)}$, based on sample $x^{(k+1)}, y^{(k+1)}$:

$$g^{(k)} = l'(w^{(k)}T x^{(k+1)} - y^{(k+1)}) x^{(k+1)}$$

where $l'$ is the derivative (or a subgradient) of $l$

on-line algorithm:

$$w^{(k+1)} = w^{(k)} - \alpha_k l'(w^{(k)}T x^{(k+1)} - y^{(k+1)}) x^{(k+1)}.$$ 

- for $l(u) = u^2$, gives the LMS (least mean-square) algorithm
- for $l(u) = |u|$, gives the sign algorithm
- $w^{(k)}T x^{(k+1)} - y^{(k+1)}$ is the prediction error
Example: Mean-absolute error minimization

Problem instance: \( n = 10, (x, y) \sim \mathcal{N}(0, \Sigma), \Sigma \) random with \( \mathbb{E}(y^2) \approx 12, \)
\( \alpha_k = 1/k \)
empirical distribution of prediction error for $w^*$ (over 1000 samples)