Optimization

Màster de Fonaments de Ciència de Dades

Gerard Gómez

Lecture IX. Stochastic optimization methods

Stochastic optimization methods

- **Stochastic optimization problems** arise in areas such as:
    - Optimization procedures where Monte Carlo simulations are run as estimates of the state or the parameters of a certain system
    - Problems where there are experimental random errors in the measurements
    - Real-time estimation and control
- **Stochastic optimization** (SO) methods are optimization methods that generate and use random variables[1].
- These methods **generalize deterministic methods** for deterministic problems.
- The **random variables** can **appear** in the formulation of the optimization problem itself, which involve **random objective functions or random constraints**.

[1] A **random variable** (aleatory variable, or stochastic variable) is a function whose values depend in some deterministic way on the set Ω of possible outcomes of a random event, $X : \Omega \to \mathbb{R}$

Stochastic optimization methods

- Stochastic optimization methods also include **methods with random iterates**, for instance: **genetic algorithms**.

- Some stochastic optimization methods, that use random iterates to solve stochastic problems, **combine both meanings** of stochastic optimization: **stochastic problems** and **stochastic methods**.

- In stochastic problems knowledge that the function values are contaminated by random "noise" leads naturally to **algorithms that use statistical tools to estimate the "true" values of the function**.

- Like deterministic optimization, there is **no single solution method** that works well for all problems. Some assumptions, such as convexity, or limits on the size of the decision and outcome spaces, are needed to make problems tractable.

Methods for stochastic optimization problems

One classification of solution methods for stochastic optimization problems is between:

- **Solution methods for single stage problems** (problems with a single time period)

  Single stage problems try to find a single, optimal decision, such as the best set of parameters for a statistical model given data.

  Single stage problems are usually solved with modified deterministic optimization methods.

- **Solution methods for multistage problems** (problems with multiple time periods)

  Multistage problems try to find an optimal sequence of decisions, such as scheduling water releases from a network of hydroelectric plants over a two year period.

  The dependence of future decisions on random outcomes makes direct modification of deterministic methods difficult in multistage problems.

  Multistage methods are more reliant on statistical approximation and strong assumptions about problem structure.

Single stage stochastic optimization

Single stage stochastic optimization is the study of optimization problems, with a random objective function or constraints, where a decision is implemented and is not corrected in further steps.

Example: Parameter selection for a statistical model. The observations are drawn from an unknown distribution, giving a random error for each observation. The goal is to select model parameters to minimize the expected error.

Single stage stochastic optimization

Formal concepts and notation.

Let $\mathcal{X}$ be the domain of all **feasible decisions** and $x \in \mathcal{X}$ a specific decision. We would like to search over $\mathcal{X}$ to find a decision that minimizes a cost function F.

Let ξ denote **random information** that is available only after the decision is made. The cost function, $F(x, \xi)$, will depend on x and ξ.

Since we cannot directly optimize $F(x, \xi)$ we instead **minimize the expected value,** $\mathbb{E}\left[F(x, \xi)\right]$[2].

The **general single stage stochastic optimization problem** becomes

$$\zeta^* = \min_{x \in \mathcal{X}} \{ f(x) = \mathbb{E}\left[F(x, \xi)\right] \}$$

For all single stage problems, we **assume that the decision space $\mathcal{X}$ is convex and the objective function $F(x, \xi)$ is convex in** x for any realization ξ.

[2] If X is a random variable, $\mathbb{E}[X]$ (also denoted by $\mathbb{E}\, X$) denotes the **expected value of** X (also called expectation, average, mean value, mean, or first moment). $\mathbb{E}[X]$ is the probability-weighted average of all possible values of X. If X is discrete, each possible value the random variable can assume is multiplied by its probability of occurring, and the resulting products are summed to produce the expected value. For a continuous random variable, an integral of the variable with respect to its probability density replaces the sum of the above definition.

Multistage stochastic optimization

Multistage stochastic optimization problems aim to find a sequence of decisions, x_t, $t = 0, ..., T$, that minimize an expected cost function. The subscript t denotes the time at which decision x_t is made.

- Usually decisions and random outcomes at time t affect the value of future decisions.
- An example would be making a move in a chess game. With a move, the player may capture one of his opponent's pieces, change his board position, and alter his possible future moves. He needs to account for these issues to select the move that maximizes his probability of winning.

Multistage stochastic optimization

- Mathematically, we can describe multistage stochastic optimization problems as an iterated expectation:

$$\zeta^* = \min_{x_0 \in \mathcal{X}_0} \mathbb{E}\left[\inf_{x_1 \in \mathcal{X}_1(x_0,\xi_1)} F_1(x_1,\xi_1) + \mathbb{E}\left[\ldots + \mathbb{E}\left[\inf_{x_T \in \mathcal{X}_T(x_{0:T-1},\xi_{1:T})} F_T(x_T,\xi_T) \right]\right]\right]$$

  - T is the number of time periods
  - $x_{0:t}$ is the collection of all decisions between 0 and t
  - ξ_t is a random outcome observable at time t
  - $\mathcal{X}_t(x_{0:t-1},\xi_{1:t})$ is a decision set that depends on all decisions and random outcomes between times 0 and t
  - $F_t(x_t,\xi_t)$ is a cost function for time period t that depends on the decision and random outcome for period t
  - The time horizon T may be either finite or infinite

- Unlike the methods for single stochastic optimization, there are no multistage solution methods that work well for all problems within a broad class, like convex problems or Markov decision processes.

Sample average approximation for single stage stochastic optimization problem

Sample average approximation (SAA) is a two steps method that uses sampling and deterministic optimization to solve

$$\zeta^* = \min_{x \in \mathcal{X}} \{ f(x) = \mathbb{E}\left[F(x, \xi) \right] \}$$

- The first step in SAA is **sampling**. While directly computing the expected cost function $\mathbb{E}\left[F(x,\xi)\right]$ is not possible for most problems, it can be approximated through Monte Carlo sampling in some situations.

  Let ξ_i, $i = 1, ..., n$ be a set of independent, identically distributed realizations of ξ, and let $F(x, \xi_i)$ be the cost function realization for ξ_i. The expected cost function is approximated by the average of the realizations:

  $$\mathbb{E}[F(x,\xi)] \approx \frac{1}{n} \sum_{i=1}^{n} F(x, \xi_i)$$

- The second step in SAA is **search**. The right hand side of the above equation is deterministic, so deterministic optimization methods can be used to solve the approximate problem:

  $$\zeta_n^* = \min_{x \in \mathcal{X}} \left\{ f_n(x) = \frac{1}{n} \sum_{i=1}^{n} F(x, \xi_i) \right\}$$

Properties of sample average approximations

Like all stochastic optimization methods, SAA relies upon a collection of random variables to produce a statistical estimate.

Most theoretical results follow directly from the the Law of Large Numbers and Central Limit Theorem due to the construction of SAA.

- Under mild regularity conditions, for any fixed x **the limiting distribution of the SAA estimate is Gaussian** with mean $f(x)$ and variance $\sigma^2(x) = \mathrm{var}(F(x,\xi))/n < \infty$:

$$\sqrt{n}\,[f_n(x) - f(x)] \to \mathcal{N}(0, \sigma^2(x)).$$

- Under more restrictive conditions, including Lipschitz continuity of $F(\cdot,\xi)$, convexity of $F(x,\xi)$, convexity of $\mathcal{X}$, and $f(x)$ having a unique optimum, a **similar result holds for the optimal values**:

$$\sqrt{n}\,[\zeta_n^* - \zeta^*] \to \mathcal{N}(0, \sigma^2(x)).$$

The limiting Gaussian distribution can be used to determine the number of samples needed to generate an ϵ-optimal solution with at least probability $1 - \alpha$.

The stochastic subgradient method. Noisy unbiased subgradients

Noisy unbiased subgradients

Let

$$f : \mathbb{R}^m \to \mathbb{R}$$

be a **convex function**.

We say that a random variable vector $\tilde{g} \in \mathbb{R}^n$ is a **noisy unbiased subgradient** of f at x if

$$\mathbb{E}\left[\tilde{g}\right] = g \in \partial f(x)$$

According to the definition of a subgradient, this means

$$f(z) \geq f(x) + \left(\mathbb{E}\left[\tilde{g}\right]\right)^T (z - x), \quad \forall z \in \mathbb{R}^m$$

Equivalently, $\tilde{g} \in \mathbb{R}^n$ is a noisy unbiased subgradient of f at x if it can be written as

$$\tilde{g} = g + v$$

where $g \in \partial f(x)$ and v **has zero mean**.

The stochastic subgradient method. Noisy subgradients

If x is also a random variable, then we say that $\tilde{g}$ is a **noisy subgradient** of f at x (which is random) if for all z

$$f(z) \geq f(x) + \left(\mathbb{E}(\tilde{g} \,|\, x)\right)^T (z - x).$$

holds almost surely (it happens with probability one or, in other words, the set of possible exceptions may be non-empty, but it has probability zero).

We can write this compactly as $\mathbb{E}(\tilde{g} \,|\, x) \in \partial f(x)$.

The noise v can represent:

- Error in computing a true subgradient,
- Error that arises in Monte Carlo evaluation of a function defined as an expected value,
- Measurement error.

Remark. The conditional expectation of a random variable X, given another random variable Y, is another random variable equal to the average of the former over all, or eventually one, possible outcomes of Y: $\mathbb{E}[(X \,|\, Y)] \equiv \mathbb{E}(X \,|\, Y)$, $\mathbb{E}[(X \,|\, Y = y)] \equiv \mathbb{E}(X \,|\, Y = y) \equiv \mathbb{E}(X \,|\, y)$

The stochastic subgradient method

The **stochastic subgradient method** is essentially the subgradient method, but **using noisy subgradients** and a more **limited set of step size rules**.

- The simplest case corresponds to **unconstrained minimization of a convex function** $f : \mathbb{R}^m \to \mathbb{R}$. In this case, the stochastic subgradient method uses the **standard update**

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

  where x^k is the k-th iterate, $\alpha_k > 0$ is the k-th step size, and $\tilde{g}^k$ is a noisy subgradient of f at x^k

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

- As in the ordinary subgradient method, we can have $f(x^k)$ **increase** during the algorithm, so we **keep track of the best point found so far**, and the associated function value

$$f_{best}^k = \min\{f(x^1), ..., f(x^k)\}.$$

The stochastic subgradient method. Convergence

We will give **convergence** results, of the stochastic subgradient method, using **step sizes that are square-summable but not summable**:

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

We will assume that:

- There is an x^* that minimizes f
- $\mathbb{E}\left[\|x^1 - x^*\|_2^2\right] \leq R^2$ for a certain $R \geq 0$
- There is a G such that $\mathbb{E}\left[\|g^k\|_2^2\right] \leq G^2$ for all k

Under these ussumptions, we will see that we have **convergence in expectation**, this is:

$$\mathbb{E}[f_{best}^k] \equiv \mathbb{E}\left[\min\{f(x^1), ..., f(x^k)\}\right] \to f^*$$

as $k \to \infty$.

We also have **convergence in probability**: for any $\epsilon > 0$

$$\lim_{k\to\infty} Prob\left(f_{best}^k \geq f^* + \epsilon\right) = 0.$$

Convergence

Proof:

We have

$$\begin{aligned}
\mathbb{E}\left[\left(\|x^{k+1}-x^*\|_2^2 \,|\, x^k\right)\right] &= \mathbb{E}\left[\left(\|x^k-\alpha_k\tilde{g}^k-x^*\|_2^2 \,|\, x^k\right)\right] \\
&= \|x^k-x^*\|_2^2 - 2\alpha_k\left[\mathbb{E}\left((\tilde{g}^k)^T(x^k-x^*) \,|\, x^k\right)\right] + \alpha_k^2\mathbb{E}\left[\left(\|\tilde{g}^k\|_2^2 \,|\, x^k\right)\right] \\
&= \|x^k-x^*\|_2^2 - 2\alpha_k\mathbb{E}\left[(\tilde{g}^k \,|\, x^k)^T(x^k-x^*)\right] + \alpha_k^2\mathbb{E}\left[\left(\|\tilde{g}^k\|_2^2 \,|\, x^k\right)\right] \\
&\leq \|x^k-x^*\|_2^2 - 2\alpha_k(f(x^k)-f^*) + \alpha_k^2\mathbb{E}\left[\left(\|\tilde{g}^k\|_2^2 \,|\, x^k\right)\right]
\end{aligned}$$

where the inequality holds almost surely, and follows from

$$\mathbb{E}\left[(\tilde{g}^k \,|\, x^k)\right] \in \partial f(x^k)$$

From the above inequality, and using the assumption $\mathbb{E}\left[\|g^k\|_2^2\right] \leq G^2$, we get

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2\right] \leq \mathbb{E}\left[\|x^k-x^*\|_2^2\right] - 2\alpha_k\left(\mathbb{E}\left[f(x^k)-f^*\right]\right) + \alpha_k^2G^2.$$

Recursively applying this inequality yields

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

Convergence (cont.)

Using

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2\right] \geq 0, \quad \mathbb{E}\left[\|x^1 - x^*\|_2^2\right] \leq R^2, \quad \sum_{i=1}^{k} \alpha_i^2 \leq \|\alpha\|_2^2,$$

we get

$$2\sum_{i=1}^{k} \alpha_i \left(\mathbb{E}\left[f(x^i) - f^*\right]\right) \leq R^2 + G^2\|\alpha\|_2^2,$$

therefore, we have

$$\min_{i=1,\dots,k} \mathbb{E}\left[f(x^i) - f^*\right] \leq \frac{R^2 + G^2\|\alpha\|_2^2}{2\sum_{i=1}^{k} \alpha_i},$$

which shows that $\min_{i=1,\dots,k} \mathbb{E}\left[f(x^i)\right]$ converges to f^*, since $\sum_{k=1}^{\infty} \alpha_k = \infty$.

Finally, we note that by Jensen's inequality and the concavity of the minimum function, we have

$$\mathbb{E}\left[f_{best}^k\right] = \mathbb{E}\left[\min_{i=1,\dots,k} f(x^i)\right] \leq \min_{i=1,\dots,k} \mathbb{E}\left[f(x^i)\right] \to f^*,$$

so $\mathbb{E}\left[f_{best}^k\right]$ also converges to f^*.

Convergence (cont.)

To show **convergence in probability**, we use Markov's inequality to obtain, for a fixed $\epsilon > 0$

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

The righthand side goes to zero as $k \to \infty$, so the lefthand side does as well.

The stochastic subgradient method. Example

We want to **minimize** the function

$$f(x) = \max_{i=1,\dots,m} (a_i^T x + b_i), \quad \text{with} \quad x, a_i \in \mathbb{R}^n \quad \text{and} \quad b_i \in \mathbb{R}.$$

We use a stochastic subgradient algorithm

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

with noisy subgradient

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

where the v^k are independent zero mean random variables.

We will take $n = 20$ variables, $m = 100$ terms, and the problem problem data a_i and b_i generated from a unit normal distribution.

The norm of the vectors a_i used is on the order of $\sqrt{20} \approx 4.5$, so $\|g\| = \|a_i\| \simeq 4.5$.

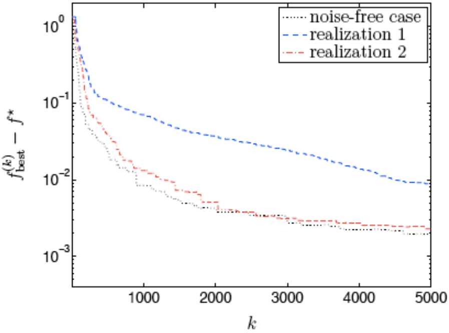
The noises v^k are independent and identically distributed normal random variables of 0 mean and variance $\sigma^2 = 0.5\, I$, and the subgradient noise is around the 25% of the true subgradient.

The stochastic subgradient method. Example (cont.)

The initial point used is the vector $x^1 = 0$, and the step rule is the square summable but not summable rule $\alpha_k = 1/k$.

The figure shows the convergence of the stochastic subgradient method for two realizations of the noisy subgradient process, together with the noise-free case for comparison.

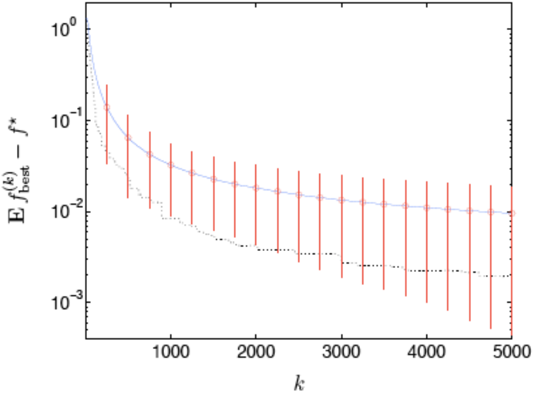
The figure also shows that convergence is only a bit slower with subgradient noise



The value of $f^* \approx 1.1$ has been obtained using linear programming.

The stochastic subgradient method. Example (cont.)

For 100 realizations of the procedure for each value of k, the the error bars show the sample mean plus and minus one standard deviation of $\mathbb{E}\left[f_{best}^{k} - f^{*}\right]$ for k in multiples of 250.



The stochastic programming problem

The **stochastic programming problem** has the form

$$\begin{cases} \text{minimize} & \mathbb{E}\left[f_0(x,w)\right] \\ \text{subject to} & \mathbb{E}\left[f_i(x,w)\right] \leq 0, \quad i = 1, \ldots, m \end{cases}$$

where $x \in \mathbb{R}^n$ is the optimization variable, and w is a random parameter.

Stochastic programming problems are used to model a variety of robust design or decision problems with **uncertain data** w.

If the $f_i(x,w)$ are convex in x for each w, the problem is a **convex stochastic programming problem**. In this case the objective and constraint functions are convex.

Variations of the stochastic programming problem. The certainty equivalent problem

Using Jensen's inequality for the convex functions f_i:

$$\mathbb{E}\left[f_i(x,w)\right] \geq f_i(x,\mathbb{E}\left[w\right])$$

we can consider the problem

$$\begin{cases} \text{minimize} & f_0(x,\mathbb{E}\left[w\right]) \\ \text{subject to} & f_i(x,\mathbb{E}\left[w\right]) \leq 0, \quad i=1,\dots,m \end{cases}$$

obtained by replacing the random variable in each function by its expected value.

- This problem is sometimes called the **certainty equivalent of the original stochastic programming problem**, even though they are equivalent only in very special cases.
- By Jensen's inequality, **the constraint set** for the uncertainty equivalent problem **is larger** than the original stochastic problem, and **its objective is smaller**.
- It follows that the optimal value of the uncertainty equivalent problem gives a **lower bound on the optimal value of the stochastic original problem**.

Other variations of the stochastic programming problem

Although the basic form of the stochastic programming problem involves only expectation or average values, some tricks can be used to capture other measures of the probability distributions of $f_i(x, w)$.

- We can replace an objective or constraint term $\mathbb{E}\left[f(x, w)\right]$ with $\mathbb{E}\left[\Phi(f(x, w))\right]$, where Φ is a convex increasing function.

  For example, with $\Phi(u) = \max(u, 0)$, we can form a constraint of the form

  $$\mathbb{E}\left[f_i(x, w)_+\right] \leq \epsilon$$

  where ϵ is a positive parameter, and $(\cdot)_+$ denotes positive part.

  Here $\mathbb{E}\left[f_i(x, w)_+\right]$ has a simple interpretation as the expected violation of the i-th constraint.

- It's also possible to combine the constraints, using a single constraint of the form

  $$\mathbb{E}\left[\max\left(f_1(x, w)_+, ..., f_1(x, w)_+\right)\right] \leq \epsilon$$

  The left hand side here can be interpreted as the expected worst violation (over all constraints).

- Unfortunately, the constraint $Prob(f_i(x, w) \leq 0) \geq \eta$ is convex only in a few special cases

Noisy subgradient of the expected function value

In order to define the **expected function value**, let $F(x, w)$

$$F : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$$

be a convex function in x for each w.

We can think of x as some kind of **optimization variable** to be chosen, w as some kind of **random parameter**

The function F gives the **cost of choosing** x **when** w **takes a particular value**

The expected function value of F is defined as

$$f(x) = \mathbb{E}\left[F(x, w)\right] = \int_{\mathbb{R}^p} F(x, w)p(w)dw$$

where p is the density of w[3]. The function f, which is deterministic and convex, gives the **average cost of choosing** x**, taking the statistical variation of** w **into account.**

Let us see how to **compute a noisy unbiased subgradient of** f **at** x.

[3] A density function $p(w)$ is a nonnegative function such that $\int_{\mathbb{R}^p} p(w)dw = 1$

Noisy subgradient of the expected function value

Except in some very special cases, we cannot easily compute $f(x)$ exactly. However, we can approximately compute f using Monte Carlo methods, if we can cheaply generate samples of w from its distribution.

- We generate M independent samples $w_1, \ldots, w_M$, and then take

  $$\hat{f}(x) = \frac{1}{M} \sum_{i=1}^{M} F(x, w_i)$$

  as our estimate of $f(x)$.

- We hope that if M is large enough, we get a good estimate. In fact, $\hat{f}(x)$ is a random variable with $\mathbb{E}\left[\hat{f}(x)\right] = f(x)$, and a variance equal to c/M, where c is the variance of $F(x, w)$.

As a summary: we cannot evaluate $f(x)$ exactly, but we can get a good approximation, with (possibly) much effort

Noisy subgradient of the expected function value

Let $G : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$ be a function that satisfies

$$G(x, w) \in \partial_x F(x, w)$$

for each x and w. In other words, $G(x, w)$ is a subgradient of F for each value of x and w.

If $F(x, w)$ is differentiable in x then we must have $G(x, w) = \nabla_x F(x, w)$.

We claim that

$$g = \mathbb{E}\left[G(x, w)\right] = \int G(x, w)p(w)dw \in \partial f(x)$$

Proof. Note that for each w and any z we have

$$F(z, w) \geq F(x, w) + G(x, w)^T(z - x)$$

since $G(x, w) \in \partial_x F(x, w)$. Multiplying this by $p(w)$, which is nonnegative, and integrating gives

$$\int F(z, w)p(w)dw \geq \int \Big(F(x, w) + G(x, w)^T(z - x)\Big)\, p(w)dw = f(x)+g^T(z-x)$$

Since the lefthand side is $f(z)$, we have shown that $g \in \partial f(x)$.

Noisy subgradient of the expected function value

Now we can **compute a noisy unbiased subgradient of f at x.**

Generate independent samples $w_1, \ldots, w_M$ and take

$$\tilde{g} = \frac{1}{M}\sum_{i=1}^{M} G(x, w_i)$$

In other words, evaluate a subgradient of f at x, for M random samples of w, and take $\tilde{g}$ to be the average.

At the same time we can also compute $\hat{f}(x)$, the Monte Carlo estimate of $f(x)$.

We have $\mathbb{E}\left[\tilde{g}\right] = \mathbb{E}\left[G(x, w)\right] = g$, which we have shown that is a subgradient of f at x.

Thus,$\tilde{g}$ **is a noisy unbiased sugradient of f at x.**

This result is independent of M.

- We can even take $M = 1$. In this case, $\tilde{g} = G(x, w_1)$. In other words, we simply generate one sample w_1 and use the subgradient of F for that value of w. In this case $\tilde{g}$ could hardly be called a good approximation of a subgradient of f, but its mean is a subgradient, so it is a valid noisy unbiased subgradient.
- On the other hand, we can take M to be large. In this case, $\tilde{g}$ is a random vector with mean g and very small variance, i.e., it is a good estimate of g, a subgradient of f at x.

Example: Expected value of piecewise linear function

Consider the problem of minimizing the expected value of a piecewise-linear convex function with random coefficients,

$$\text{minimize } f(x) = \mathbb{E}\left[\max_{i=1,\dots,m}(a_i^T + b_i)\right]$$

with variable $x \in \mathbb{R}^n$. The data $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ are random with some given distribution.

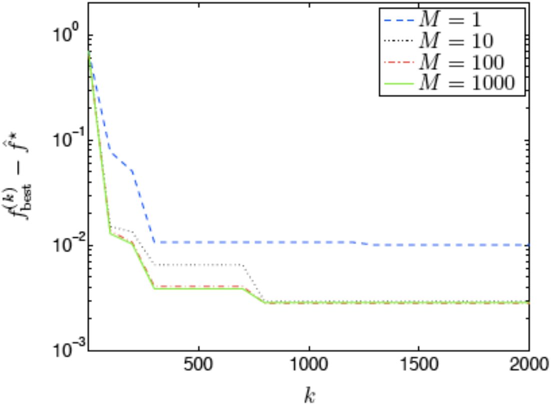
We can compute an (unbiased) approximation of $f(x)$, and a noisy unbiased subgradient $g \in \partial f(x)$, using Monte Carlo methods.

We consider a problem instance with $n = 20$ variables and $m = 100$ terms and assumed that

$$a_i \sim \mathcal{N}(\bar{a}_i, 5I) \qquad b_i \sim \mathcal{N}(\bar{b}_i, 5I)$$

The mean values $\bar{a}_i$ and $\bar{b}_i$ are generated from unit normal distributions. We take $x^1 = 0$ as the starting point and use the step size rule $\alpha_k = 1/k$.

Example: Expected value of piecewise linear function



The value of $f_{best}^k - f^*$ versus iteration number k, for the stochastic subgradient method with step size rule $\alpha_k = 1/k$. The noisy subgradients have been evaluated using M = 1, M = 10, M = 100, and M = 1000.

Example: Adaptive signal processing

Assume that $(x, y) \in \mathbb{R}^n \times \mathbb{R}$ have some joint distribution.

The goal is to **find a weight vector** $w \in \mathbb{R}^n$ **for which** $w^T x$ **is a good estimator of** y, this is: w must minimize

$$J(w) = \mathbb{E}\left[L(w^T x - y)\right]$$

where $L : \mathbb{R} \to \mathbb{R}$ is a convex loss function.

For example:

- If $L(u) = u^2$, J is the mean-square error. This is the usual loss function in signal processing.
- If $L(u) = |u|$, J is the mean-absolute error.
- If $L(u) = \max\{|u| - 1, 0\}$, L is a loss function with dead zone.

Since J is convex, minimizing J over w is a convex optimization problem.

Example: Adaptive signal processing

We consider a setting where we do not know the distribution of (x, y).

- At each step (which, for example, might correspond to time), we are given a sample (x^i, y^i) from the distribution.
- After k steps, we could use the k samples to estimate the distribution; for example, we could choose w to minimize the average loss under the empirical distribution.

  This approach requires that we store all past samples, and we need to solve a large problem each time a new sample is added.
- Instead we can **use the stochastic subgradient method.** We carry out a stochastic subgradient step each time a new sample is added.

Example: Adaptive signal processing

- Suppose we are given a new sample (x^{k+1}, y^{k+1}), and the current weight value is w^k.

- Form the following noisy unbiased subgradient of J:

$$L'\left((w^k)^T x^{k+1} - y^{k+1}\right) x^{k+1}$$

  where L' is the derivative of L. If L is nondifferentiable, substitute any subgradient of L in place of L'.

- Then, the algorithm is:
  - When (x^{k+1}, y^{k+1}) becomes available,
  - Update w^k as follows:

$$w^{k+1} = w^k - \alpha_k L'\left((w^k)^T x^{k+1} - y^{k+1}\right) x^{k+1}$$

    If $L(u) = u^2$, this is the usual update in signal processing.
  - Use (for example) $\alpha_k = 1/k$.

- Note that $(w^k)^T x^{k+1} - y^{k+1}$ is the prediction error for the $k+1$ sample, using the weight from step k.

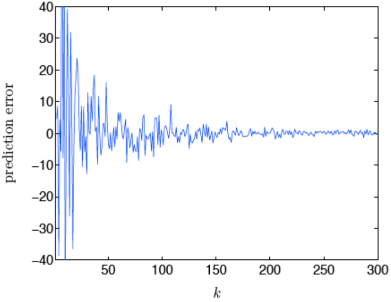
Example: Adaptive signal processing

Illustration of the method using $n = 10$, $(x, y) \sim \mathcal{N}(0, \sigma)$, where σ is chosen randomly, and $L(u) = |u|$.

The update, that in adaptive signal processing is called the **sign algorithm**, is

$$w^{k+1} = w^k - \alpha_k \operatorname{sign}\left((w^k)^T x^{k+1} - y^{k+1}\right) x^{k+1}.$$

We use $\alpha_k = 1/k$, and compute w^* using the above subgradient algorithm



Behaviour of the prediction error $(w^k)^T x^{k+1} - y^{k+1}$ for the first 300