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

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:

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

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

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

- almost sure convergence:

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

  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

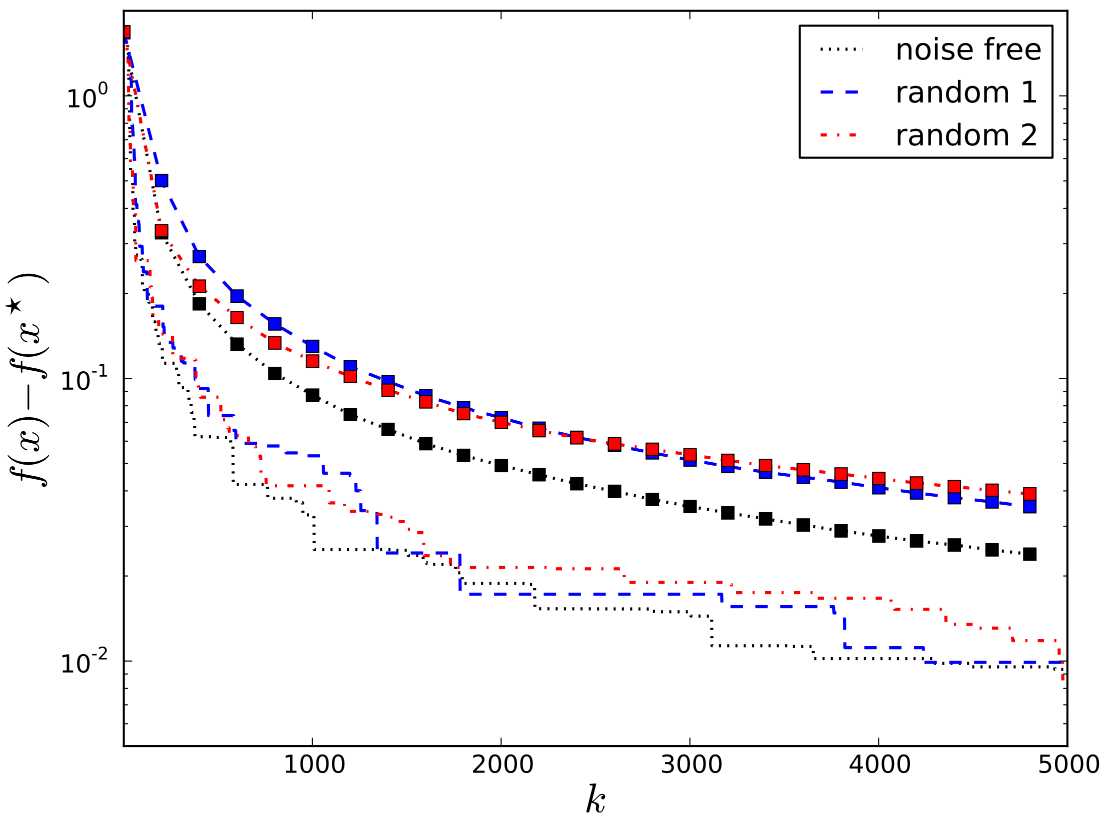
$$\text{minimize} \quad f(x) = \max_{i=1,\ldots,m}(a_i^T x + b_i)$$

we use stochastic subgradient algorithm with noisy subgradient

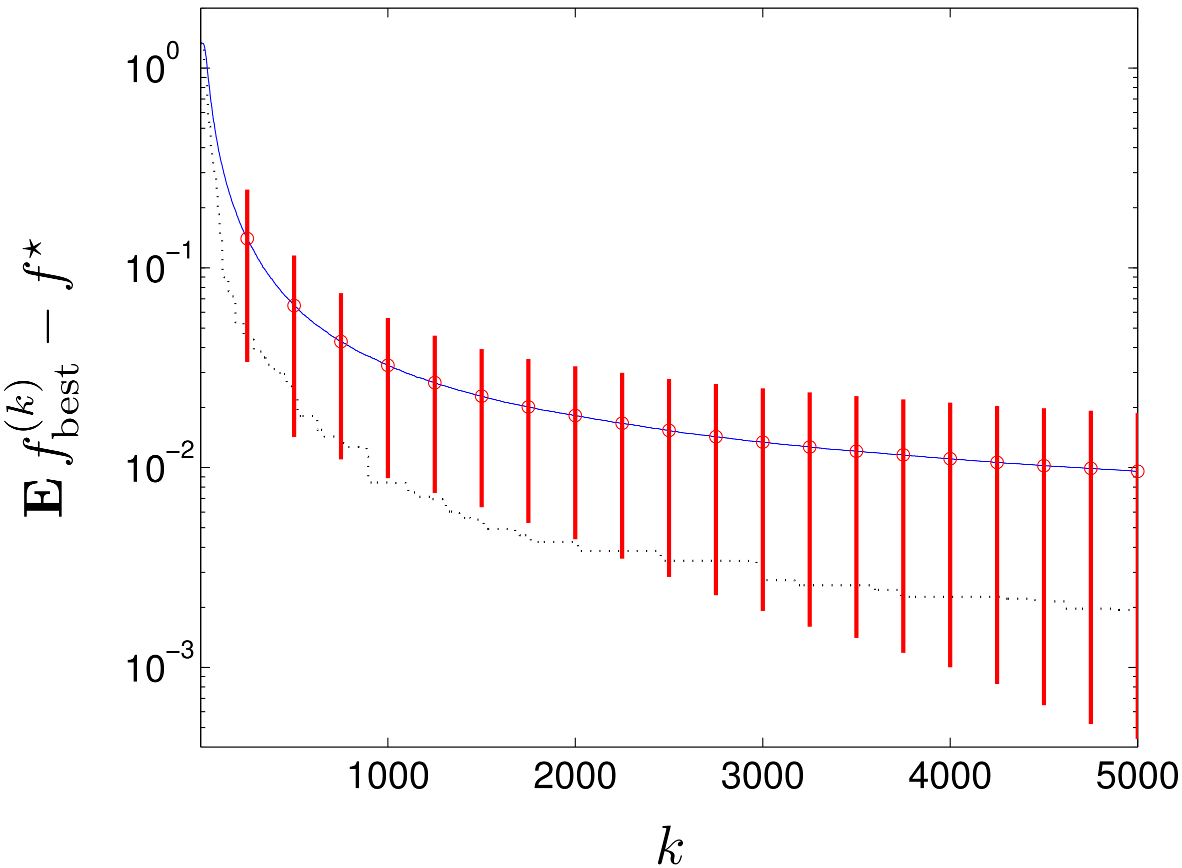
$$\tilde{g}^{(k)} = g^{(k)} + v^{(k)}, \qquad g^{(k)} \in \partial f(x^{(k)})$$

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

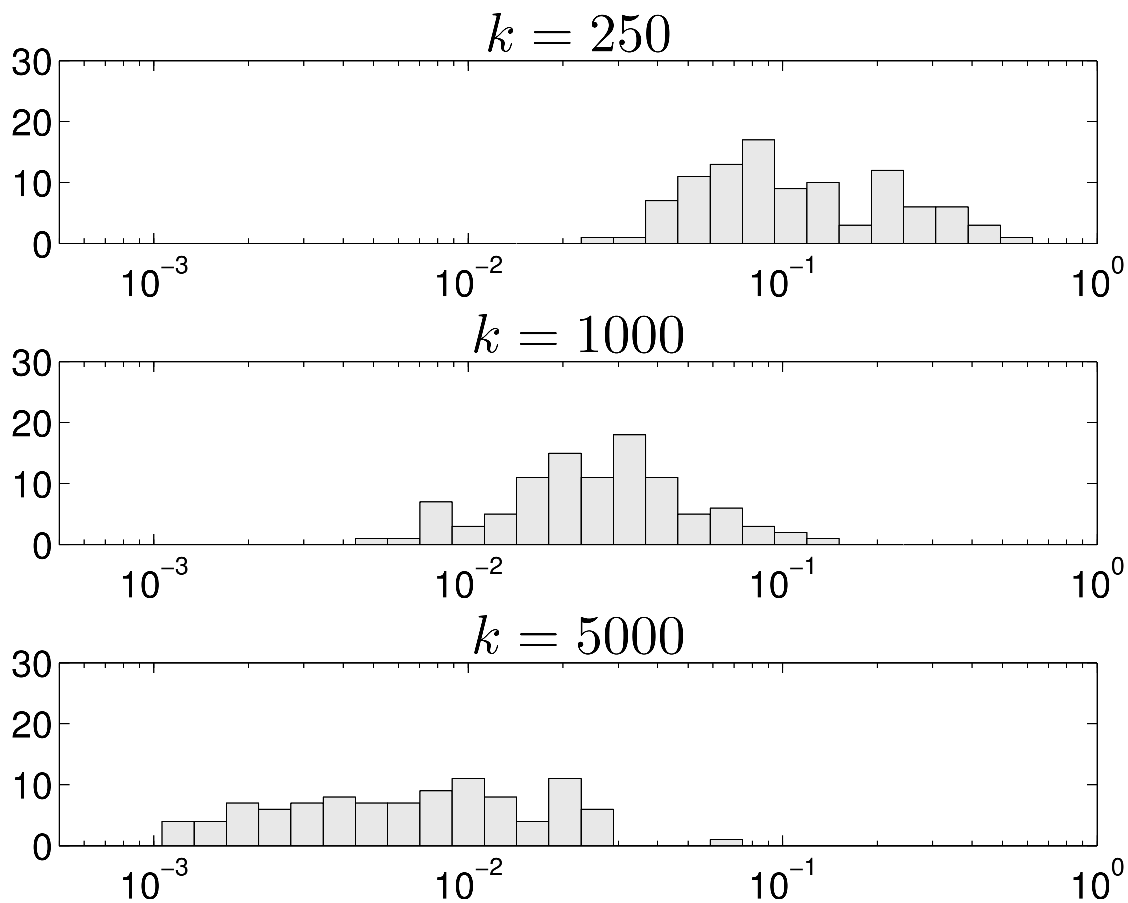
problem instance: $n = 20$ variables, $m = 100$ terms, $f^\star \approx 1.1$, $\alpha_k = 1/k$
$v^{(k)}$ are IID $\mathcal{N}(0, 0.5I)$ (25% noise since $\|g\| \approx 4.5$)



average and one std. dev. for $f_{\text{best}}^{(k)} - f^\star$ over 100 realizations



empirical distributions of $f_{\rm best}^{(k)} - f^\star$ at $k = 250$, $k = 1000$, and $k = 5000$



Stochastic programming

$$\begin{array}{ll} \text{minimize} & \mathbf{E}\, f_0(x,\omega) \\ \text{subject to} & \mathbf{E}\, f_i(x,\omega) \leq 0, \quad i=1,\ldots,m \end{array}$$

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

'certainty-equivalent' problem

$$\begin{array}{ll} \text{minimize} & f_0(x,\mathbf{E}\,\omega) \\ \text{subject to} & f_i(x,\mathbf{E}\,\omega) \leq 0, \quad i=1,\ldots,m \end{array}$$

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

Variations

- in place of $\mathbf{E}\, f_i(x, \omega) \leq 0$ (constraint holds in expectation) can use
  - $\mathbf{E}\, f_i(x, \omega)_+ \leq \epsilon$ (LHS is expected violation)
  - $\mathbf{E}\left(\max_i f_i(x, \omega)_+\right) \leq \epsilon$ (LHS is expected worst violation)
- unfortunately, *chance constraint* $\mathbf{Prob}(f_i(x, \omega) \leq 0) \geq \eta$ 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 $G(x, w) \in \partial_x F(x, w)$

- $f(x) = \mathbf{E}\, F(x, w) = \int F(x, w) p(w)\, dw$ is convex
- a subgradient of f at x is

$$g = \mathbf{E}\, G(x, w) = \int G(x, w) p(w)\, dw \in \partial f(x)$$

- 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

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

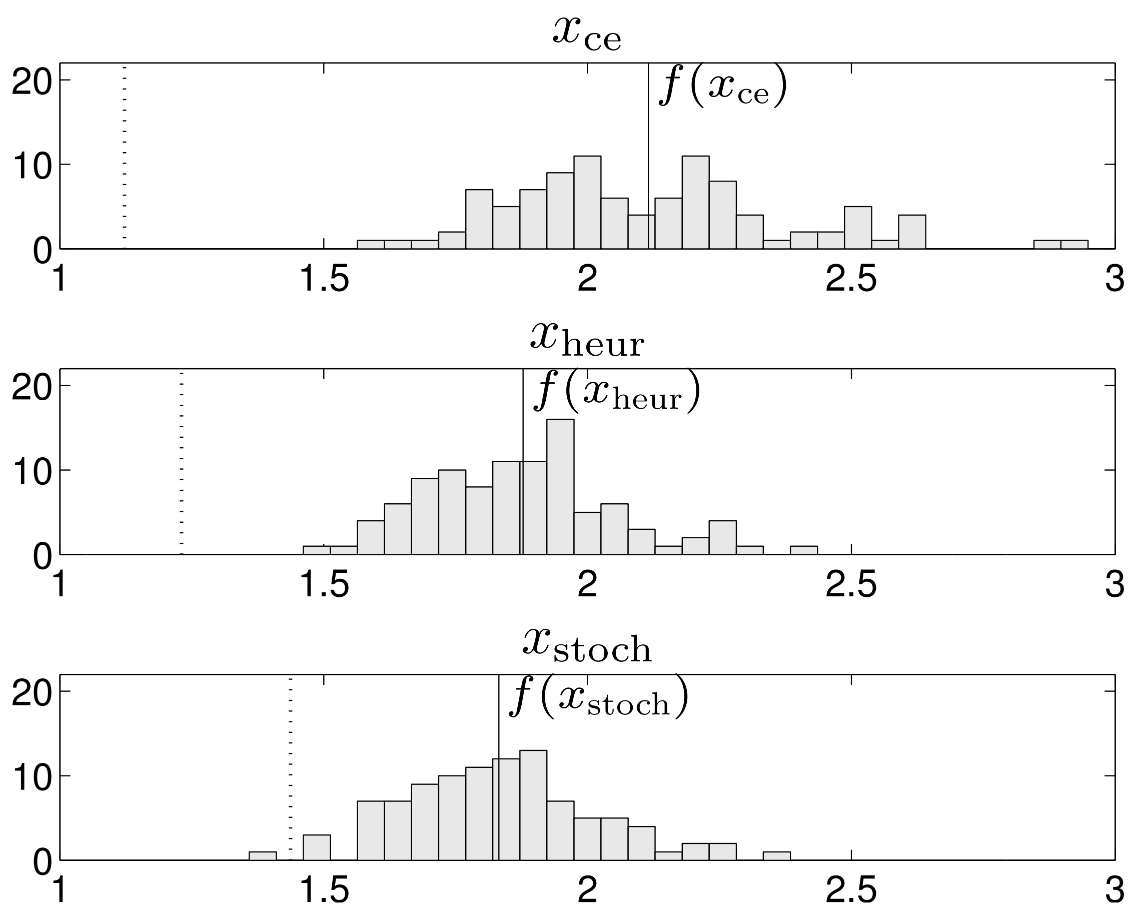
where a_i and b_i are random

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

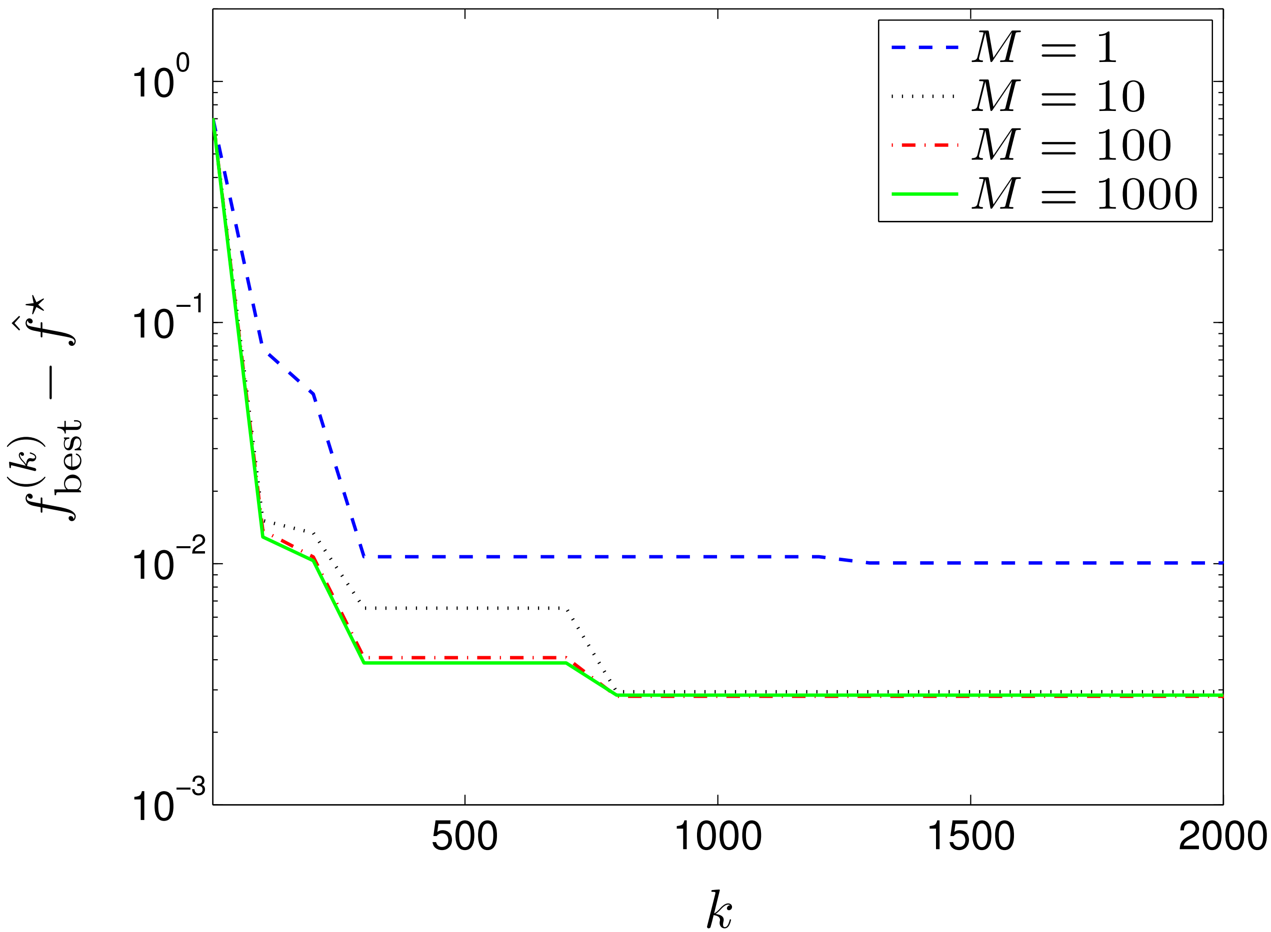
compare to:

- certainty equivalent: minimize $f_{\text{ce}}(x) = \max_{i=1,\ldots,m}(\mathbf{E}\, a_i^T x + \mathbf{E}\, b_i)$
- heuristic: minimize $f_{\text{heur}}(x) = \max_{i=1,\ldots,m}(\mathbf{E}\, a_i^T x + \mathbf{E}\, b_i + \lambda \|x\|_2)$

problem instance: $n = 20$, $m = 100$, $a_i \sim \mathcal{N}(\bar{a}_i, 5I)$, $b \sim \mathcal{N}(\bar{b}, 5I)$, $\|a_i\|_2 \approx 5$, $\|b\|_2 \approx 10$, $x_{\rm stoch}$ computed using $M = 100$



$f^\star \approx 1.34$ estimated by running the method with $M = 1000$ for long time



On-line learning and adaptive signal processing

- $(x, y) \in \mathbf{R}^n \times \mathbf{R}$ have some joint distribution
- find weight vector $w \in \mathbf{R}^n$ for which $w^T x$ is a good estimator of y
- choose w to minimize expected value of a convex *loss function* l

$$J(w) = \mathbf{E}\, l(w^T x - y)$$

  - $l(u) = u^2$: 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

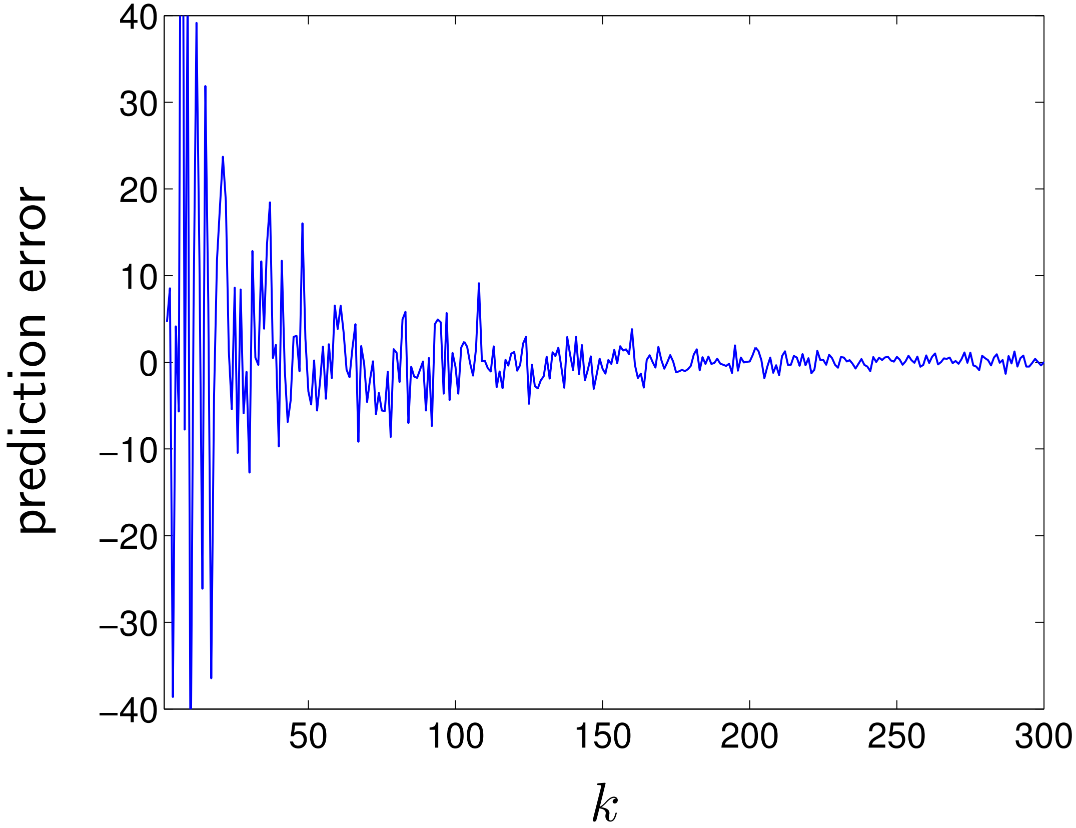
on-line algorithm:

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

- 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: $n = 10$, $(x, y) \sim \mathcal{N}(0, \Sigma)$, Σ random with $\mathbf{E}(y^2) \approx 12$, $\alpha_k = 1/k$



empirical distribution of prediction error for $w^\star$ (over 1000 samples)