Chapter 6-7: stochastic algorithms for large scale problems

Edouard Pauwels

Statistics and optimization in high dimensions
M2RI, Toulouse 3 Paul Sabatier

Different estimators

$\mathbb{X} \in \mathbb{R}^{n \times d}$, $Y \in \mathbb{R}^n$ (random).

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \|\mathbb{X}\theta - Y\|_2^2$$

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \|\mathbb{X}\theta - Y\|_2^2, \qquad \text{s.t.} \qquad \|\theta\|_1 \leq 1$$

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \|\mathbb{X}\theta - Y\|_2^2, \qquad \text{s.t.} \qquad \|\theta\|_0 \leq k$$

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \|\mathbb{X}\theta - Y\|_2^2 + \lambda\|\theta\|_0$$

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \|\mathbb{X}\theta - Y\|_2^2 + \lambda\|\theta\|_1$$

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \|\theta\|_1, \qquad \text{s.t.} \qquad \mathbb{X}\theta = Y.$$

- How to deal with large values of n (possibly infinite)?
- Can we reduce the cost of treating large d.

Where have we been so far.

- $\|\cdot\|_0$: hard to handle computationally.
- ℓ_1 norm estimators are solutions to conic programs.
- General purpose solvers (interior point methods), hardly apply to large instances.
- Dedicacted first order methods, cheap iterations.

Plan for today: stochastic algorithms to treat large n or large d.

- Stochastic approximation and Robbins-Monro algorithm.
- Prototype algorithm, ODE method, convergence rate analysis.
- Block coordinate methods, convergence rate analysis.
- General conclusion, **I expect your feedback**.

Sources are diverse, see the lecture notes.

Plan

Motivation for large n

The Lasso estimator is given as follows:

$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \|\mathbb{X}\theta - Y\|^2 + \lambda \|\theta\|_1$$

$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} (x_i^T \theta - y_i)^2 + \lambda \|\theta\|_1,$$

General model

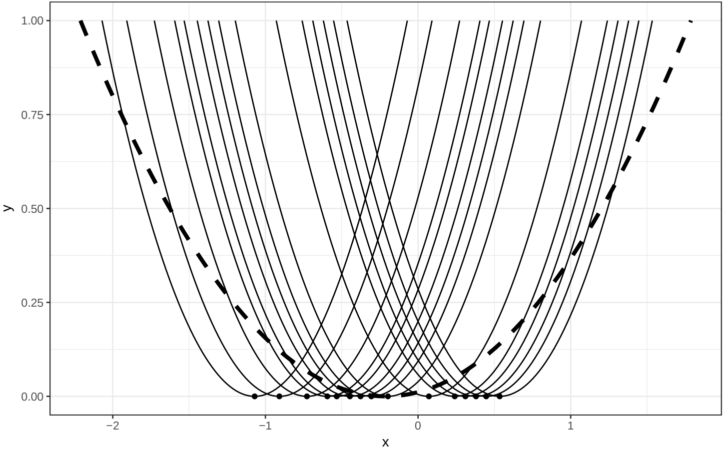
$$\min_{x \in \mathbb{R}^p} \quad F(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + g(x). \tag{1}$$

where f_i and g are convex lower semicontinuous convex functions.

Sum rule: $\partial F = \sum_{i=1}^{n} \partial f_i + \partial g$

Redundancy: if $f_i = f$ for all $i = 1, \ldots, n$, the sum is not needed, only one term.

$$\min_{x \in \mathbb{R}} F(x) := \frac{1}{n} \sum_{i=1}^{n} (x - x_i)^2$$



Intuition for stochastic approximation

Let I be uniform over $\{1, \ldots, n\}$

$$F \colon x \mapsto \mathbb{E}\left[f_I(x)\right] + g(x),$$

Stochastic approximation: main algorithmic step, for any $x \in \mathbb{R}^d$,

- Sample i uniformly at random in $\{1, \ldots, n\}$.
- Perform an algorithmic step using only the value of $f_i(x)$ and $\nabla f_i(x)$ or eventually $v \in \partial f_i(x)$

Unbiased estimates of the (sub)gradient.

- If for each value of I, f_I is $\mathcal{C}^1$, we have for any $x \in \mathbb{R}^d$,

$$\mathbb{E}\left[\nabla f_I(x)\right] = \nabla \mathbb{E}\left[f_I(x)\right] = \nabla F(x)$$

- Let v_I be a random variable such that $v_I \in \partial f_I(x)$ almost surely, F is convex and

$$\mathbb{E}\left[v_I\right] \in \partial \mathbb{E}\left[f_I(x)\right] = \partial F(x).$$

Plan

Robbins-Monro algorithm

Let $h\colon \mathbb{R}^p \mapsto \mathbb{R}^p$ be Lipschitz, we seek a zero of h, noisy unbiased estimates of h.

Robins-Monro: $(X_k)_{k\in\mathbb{N}}$ is a sequence of random variables such that for any $k \in \mathbb{N}$

$$X_{k+1} = X_k + \alpha_k \left(h(X_k) + M_{k+1}\right) \tag{2}$$

where

- $\left(\alpha_k\right)_{k\in\mathbb{N}}$ is a sequence of positive step sizes satisfying

$$\sum_{i=1}^{n} \alpha_k = +\infty \qquad \sum_{i=1}^{n} \alpha_k^2 < +\infty$$

- $(M_k)_{k\in\mathbb{N}}$, martingale difference sequence with respect to the increasing σ-fields

$$\mathcal{F}_k = \sigma(X_m, M_m, m \leq k) = \sigma(X_0, M_1, \ldots, M_k).$$

  $\mathbb{E}\left[M_{k+1}|\mathcal{F}_k\right] = 0$, for all $k \in \mathbb{N}$.

- In addition, we assume that there exists a positive constant C such that

$$\sup_{k\in\mathbb{N}} \mathbb{E}\left[\|M_{k+1}\|_2^2|\mathcal{F}_k\right] \leq C.$$

Martingale convergence theorem: $\sum_{k=0}^{+\infty} \mathbb{E}\left[\alpha_k^2 \|M_{k+1}\|^2 | \mathcal{F}_k\right]$ is finite. Hence

$$\sum_{k=0}^{K} \alpha_k M_{k+1}$$

is a zero mean martingale with square summable increments. It converges to a square integrable random varible M in $\mathbb{R}^p$, almost surely and in L^2 (Durret Section 5.4).

Vanishing step size: In addition to wash out noise, we obtain trajectories close to the ODE

$$\dot{x} = h(x)$$

Plan

1. Introduction to stochastic approximation
2. Robbins-Monro algorithm
3. Convergence analysis
4. Block coordinate algorithms

The ODE method

Choose $h = -\nabla F(x)$ assuming that F has Lispchitz gradient. The following result is due to Michel Benaim.

Theorem

Conditioning on boundedness of $\{X_k\}_{k\in\mathbb{N}}$, almost surely, the (random) set of accumulation point of the sequence is compact connected and invariant by the flow generated by the continuous time limit:

$$\dot{x} = h(x).$$

Consequence: let $\bar{x}$ be an accumulation point, the unique solution $x\colon t \mapsto \mathbb{R}^p$ to $\dot{x} = -\nabla F(x)$, $x(0) = \bar{x}$ remains bounded for all $t \in \mathbb{R}$.

Corollary

If F is convex, $\mathcal{C}^1$ with Lipschitz gradient, and attains its minimum, setting $h = -\nabla F$, conditioning on the event that $\sup_{k\in\mathbb{N}} \|X_k\|$ is finite, almost surely, all the accumulation points of X_k are critical points of F.

Non asymptotic rates for stochastic subgradient

Proposition

Consider the problem

$$\min_{x \in \mathbb{R}^d} F(x) \;:=\; \frac{1}{n} \sum_{i=1}^{n} f_i(x),$$

where each f_i *is convex and L-Lipschitz. Choose* $x_0 \in \mathbb{R}$ *and a sequence of random variables* $(i_k)_{k \in \mathbb{N}}$ *independently identically distributed uniformly on* $\{1, \ldots, n\}$ *and a sequence of positive step sizes* $(\alpha_k)_{k \in \mathbb{N}}$*. Consider the recursion*

$$x_{k+1} = x_k - \alpha_k v_k \tag{3}$$

$$v_k \in \partial f_{i_k}(x_k) \tag{4}$$

Then for all $K \in \mathbb{N}$, $K \geq 1$

$$\mathbb{E}\left[F(\bar{x}_K) - F^*\right] \leq \frac{\|x_0 - x^*\|_2^2 + L^2 \sum_{k=0}^{K} \alpha_k^2}{2 \sum_{k=0}^{K} \alpha_k}$$

where $\bar{x}_K = \frac{\sum_{k=0}^{K} \alpha_k x_k}{\sum_{k=0}^{K} \alpha_k}$.

Consequences

Corollary

Under the same hypotheses, we have the following

- *If $\alpha_k = \alpha$ is constant, we have*

$$\mathbb{E}\left[F(\bar{x}_k) - F^*\right] \leq \frac{\|x_0 - x^*\|^2}{2(k+1)\alpha} + \frac{L^2\alpha}{2}.$$

- *In particular, choosing $\alpha_i = \frac{\|x_0 - x^*\|/L}{\sqrt{k+1}}$, we have*

$$\mathbb{E}\left[F(\bar{x}_k) - F^*\right] \leq \frac{\|x_0 - x^*\| L}{\sqrt{k+1}}.$$

- *Choosing $\alpha_k = \|x_0 - x^*\|/(L\sqrt{k})$ for all k, we obtain for all k*

$$\mathbb{E}\left[F(\bar{x}_k) - F^*\right] = O\left(\frac{\|x_0 - x^*\|_2 L(1 + \log(k))}{\sqrt{k}}\right).$$

Exercise

For the last point, what can you say if F is strongly convex?

Proposition

Consider the problem

$$\min_{x\in\mathbb{R}^d} F(x) \; := \; \frac{1}{n}\sum_{i=1}^{n} f_i(x) + g(x)$$

where each f_i is convex with L-Lipschitz gradient and g is convex. Choose $x_0 \in \mathbb{R}$ and a sequence of random variables $(i_k)_{k\in\mathbb{N}}$ independently identically distributed uniformly on $\{1,\ldots,n\}$ and a sequence of positive step sizes $(\alpha_k)_{k\in\mathbb{N}}$. Consider the recursion

$$x_{k+1} = \mathrm{prox}_{\alpha_k g/L}\left(x_k - \alpha_k/L \nabla f_{i_k}(x_k)\right). \tag{5}$$

Assume the following

- $0 < \alpha_k \leq 1$, *for all* $k \in \mathbb{N}$.
- f_i *and* g *are* G*-Lipschitz for all* $i = 1, \ldots, n$;

Then for all $K \in \mathbb{N}$, $K \geq 1$, *setting* $\bar{x}_K = \frac{\sum_{k=0}^K \alpha_k x_k}{\sum_{k=0}^K \alpha_k}$.

$$\mathbb{E}\left[F(\bar{x}_K) - F^*\right] \leq \frac{L\|x_0 - x^*\|_2^2 + \frac{2G^2}{L}\sum_{k=0}^K \alpha_k^2}{2\sum_{k=0}^K \alpha_k}$$

Corollary

- *If* $\alpha_k = \alpha$ *is constant, we have for all* $k \geq 1$

$$F(\bar{x}_k) - F^* \leq \frac{L\|x_0 - x^*\|^2}{2(k+1)\alpha} + \frac{G^2\alpha}{L}.$$

- *In particular, choosing* $\alpha_i = \frac{1}{\sqrt{2k+2}}$, *for* $i = 1\dots, k$, *for some* $k \in \mathbb{N}$, *we have*

$$F(\bar{x}_k) - F^* \leq \frac{L\|x_0 - x^*\|_2^2 + \frac{G^2}{L}}{\sqrt{2k+2}}.$$

- *Choosing* $\alpha_k = 1/\sqrt{2k+2}$ *for all* k, *we obtain for all* k

$$F(x_k) - F^* = O\left(\frac{L\|x_0 - x^*\|_2^2 + \frac{G^2}{L}\log(k)}{\sqrt{2k+2}}\right).$$

Optimality of these rates

- $O(1/\sqrt{k})$ are optimal rates for optimization based on stochastic oracles.
- Smoothness does not improve.
- Strong convexity leads to $O(1/k)$.
- Linear rates can be achieved using variance reduction techniques for finite sums under strong convexity.

Minimizing the population risk

Minimize functions of the form

$$x \mapsto \mathbb{E}_Z \left[f(x, Z) \right]$$

where x are model parameters and Z is a population random variable.

Example: input output pair (X, Y) of a regression problem, minimize over a parametric regression function class $\mathcal{F}$.

$$R(f) = \mathbb{E}\left[(f(X) - Y)^2 \right] = \int_{\mathcal{X} \times \mathcal{Y}} (f(x) - y))^2 P(dx, dy).$$

Single pass: given $(x_i, y_i)_{i=1}^n$, one pass of a stochastic algorithm, amount to perform n steps of the same algorithm on the population risk.

Plan

Motivation for large d

$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \|\mathbb{X}\theta - Y\|^2 + \lambda \|\theta\|_1.$$

The cost of one proximal gradient step is of the order of d^2.

Idea: Update only subsets of the coordinates to reduce the cost.

Does this work

- For smooth convex functions?
- For nonsmooth convex functions?
- For the Lasso problem?

Block proximal gradient algorithm

We consider optimization problems of the form

$$\min_{x \in \mathbb{R}^p} F(x) = f(x) + \sum_{i=1}^{p} g_i(x_i),$$

where $f : \mathbb{R}^p \mapsto \mathbb{R}$ has L-Lipschitz gradient and $g_i : \mathbb{R} \mapsto \mathbb{R}$ are convex lower semicontinuous univariate functions.

Let $e_1, \ldots, e_p$ be the elements of the canonical basis. Given a sequence of coordinate indices $(i_k)_{k \in \mathbb{N}}$, starting at $x_0 \in \mathbb{R}^p$

$$x_{k+1} = \arg\min_{y = x_k + t e_{i_k}} f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2} \|y - x_k\|_2^2 + g_{i_k}(y)$$

Assumption (Coercivity)

The sublevelset $\{y \in \mathbb{R}^p,\ F(y) \leq F(x_0)\}$ is compact, for any $y \in \mathbb{R}^p$ such that $F(y) \leq F(x_0)$, $\|y - x^*\|_2 \leq R$.

Technical Lemma

Lemma

Let $(A_k)_{k\in\mathbb{N}}$ *be a sequence of positive real numbers and* $\gamma > 0$ *be such that*

$$A_k - A_{k+1} \geq \gamma A_k^2$$

then for all $k \in \mathbb{N}$, $A_k \leq \left(\frac{1}{A_0} + \gamma k\right)^{-1}$.

Proposition (Nesterov 2012)

Consider the problem

$$\min_{x \in \mathbb{R}^p} f(x)$$

where $f : \mathbb{R}^p \mapsto \mathbb{R}$ *is convex differentiable with L-Lipschitz gradient. Choose* $x_0 \in \mathbb{R}$ *and a sequence of random variables* $(i_k)_{k \in \mathbb{N}}$ *independently identically distributed uniformly on* $\{1, \dots, p\}$ *and a sequence of positive step sizes. Consider the recursion*

$$x_{k+1} = x_k - \frac{1}{L} \nabla_{i_k} f(x_k) \tag{6}$$

Then for all $k \in \mathbb{N}$, $k \geq 1$

$$\mathbb{E}\left[f(x_k) - f^*\right] \leq \frac{2pLR^2}{k}.$$

Random block proximal gradient descent

Proposition (Richtárik, Takác 2014)

Consider the problem

$$\min_{x \in \mathbb{R}^d} F(x) \; := \; f(x) + \sum_{i=1}^{p} g_i(x)$$

where $f : \mathbb{R}^p \mapsto \mathbb{R}$ *is convex differentiable with L-Lipschitz gradient, each* $g_i : \mathbb{R}^p \mapsto \mathbb{R}$ *is convex and lower semicontinuous and only depends on coordinate* i*. Choose* $x_0 \in \mathbb{R}$ *and a sequence of random variables* $(i_k)_{k \in \mathbb{N}}$ *independently identically distributed uniformly on* $\{1, \dots, p\}$ *and a sequence of positive step sizes. Consider the recursion*

$$\begin{aligned} x_{k+1} &= \arg\min_y f(x_k) + \langle \nabla_{i_k} f(x_k), y - x_k \rangle + \frac{L}{2}\|y - x_k\|_2^2 + g_{i_k}(y) && (7) \\ &= \mathrm{prox}_{g_{i_k}/L}\left(x_k - \frac{1}{L}\nabla_{i_k} f(x_k) \right). && (8) \end{aligned}$$

Set $C = \max\left\{LR^2, F(x_0) - F^*\right\}$*, we have, for all* $k \geq 1$*,*

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \frac{2pC}{k}.$$

Deterministic block gradient descent

Proposition

Consider the problem

$$\min_{x\in\mathbb{R}^d} f(x)$$

where $f: \mathbb{R}^p \mapsto \mathbb{R}$ *is convex differentiable with L-Lipschitz gradient. Choose* $x_0 \in \mathbb{R}$, *and consider the recursion*

$$x_{k+1} = x_k - \frac{1}{L}\nabla_{i_k} f(x_k) \tag{9}$$

where i_k *is the largest block of* $\nabla f(x_k)$ *in Euclidean norm. Then for all* $k \in \mathbb{N}$, $k \geq 1$

$$f(x_k) - f^* \leq \frac{2pLR^2}{k}.$$

Similar for proximal variant.

Exercise

In the same setting as the block gradient descent method, consider the update

$$x_{k+1} \in \arg\min \quad f(y), \qquad \text{s.t.} \qquad y = x_k + te_{i_k},\ t \in \mathbb{R}.$$

Can you prove a convergence rate for this method?

Comment on complexity for quadratic losses

Lasso estimator

$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \|\mathbb{X}\theta - Y\|^2 + \lambda \|\theta\|_1.$$

- Full proximal gradient step costs $O(d^2)$.
- Given $\theta \in \mathbb{R}^d$ and $\beta = \mathbb{X}^T(\mathbb{X}\theta - Y)$, choosing $\tilde{\theta}$ such that $\|\theta - \tilde{\theta}\|_0$, computing $\mathbb{X}^T(\mathbb{X}\tilde{\theta} - Y)$ given β costs only $O(d)$.

Consequences:

- d steps of random block method have roughly the same cost as one step of the full method.
- The computational overhead for deterministic rules is affordable.

Conclusion for randomized methods

- Stochastic Gradient Descent (SGD) is at the heart of machine learning methods beyond convex optimization (deep learning . . .).
- Block decomposition methods can be beneficial even for small d.
- Most state of the art method used randomized algorithms.

Conclusion

Sparse least squares problem, a running example to illustrate:

- Statistical efficiency issues in high dimension and their resolution
- Computational complexity barriers in high dimensional estimation.
- All purpose solvers for conic programming
- First order methods and composite optimization
- Randomized methods to treat high dimensionality issues from a computational view point.

Lecture notes: available at

`https://www.math.univ-toulouse.fr/~epauwels/M2RI/index.html`

Feedback form: please rate the course.