Stochastic Approximation

Francis Bach, Aymeric Dieuleveut, Alain Durmus, Eric Moulines

Ecole Polytechnique, Centre de Mathematiques Appliquees

July 20, 2021

Context

Machine learning for "big data"

- Large-scale machine learning: large d, large n
  - d: dimension of each observation (input)
  - n : number of observations
- Examples: computer vision, bioinformatics, advertising
- Ideal running-time complexity: $O(dn)$

Context

Machine learning for "big data"

- Large-scale machine learning: large d, large n
  - d: dimension of each observation (input)
  - n : number of observations
- Examples: computer vision, bioinformatics, advertising
- Ideal running-time complexity: $O(dn)$
- Going back to simple methods
  - Stochastic gradient methods (Robbins, Monro, 1951)

Context

Machine learning for "big data"

- Large-scale machine learning: large d, large n
  - d: dimension of each observation (input)
  - n : number of observations
- Examples: computer vision, bioinformatics, advertising
- Ideal running-time complexity: $O(dn)$
- Going back to simple methods
  - Stochastic gradient methods (Robbins, Monro, 1951)
  - Mixing statistics and optimization

Supervised machine learning

- Data: n observations $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \ldots, n$, i.i.d.
- Prediction as a linear function $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathbb{R}^d$
- (regularized) empirical risk minimization: find $\hat{\theta}$ solution of

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} \ell\big(Y_i, \langle \theta, \Phi(X_i) \rangle\big) \quad + \quad \mu \Omega(\theta)$$

convex data fitting term + regularizer

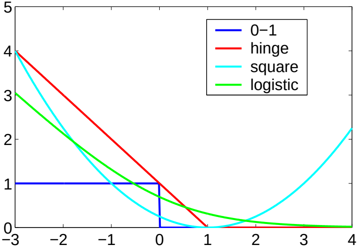
Usual losses

- Regression: $y \in \mathbb{R}$, prediction $\phi_\theta(x) = \langle \theta, \Phi(x) \rangle$
  - quadratic loss $\ell(y, \langle \theta, \Phi(x) \rangle) = \frac{1}{2}(y - \langle \theta, \Phi(x) \rangle)^2$

Usual losses

- Regression: $y \in \mathbb{R}$, prediction $\phi_\theta(x) = \langle \theta, \Phi(x) \rangle$
  - quadratic loss $\ell(y, \langle \theta, \Phi(x) \rangle) = \frac{1}{2}(y - \langle \theta, \Phi(x) \rangle)^2$
- Classification: $y \in \{-1, 1\}$, prediction $\phi_\theta(x) = \mathrm{sign}(\langle \theta, \Phi(x) \rangle)$
  - $0 - 1$ loss: $\ell(y, \langle \theta, \Phi(x) \rangle) = \mathbf{1}_{\{y \cdot \langle \theta, \Phi(x) \rangle < 0\}}$.
  - convex losses

Convex loss



- Support vector machine (hinge loss)

$$\ell(Y, \langle\theta, \Phi(x)\rangle) = \max\{1 - Y\langle\theta, \Phi(x)\rangle, 0\}$$

- Logistic regression:

$$\ell(Y, \langle\theta, \Phi(x)\rangle) = \log(1 + \exp(-Y\langle\theta, \Phi(x)\rangle))$$

- Least-squares regression

$$\ell(Y, \langle\theta, \Phi(x)\rangle) = \frac{1}{2}(Y - \langle\theta, \Phi(x)\rangle)^2$$

Usual regularizers

- Main goal: avoid overfitting
- (squared) Euclidean norm: $\|\theta\|_2^2 = \sum_{j=1}^d |\theta_j|^2$
- Sparsity-inducing norms
  - LASSO : ℓ_1-norm $\|\theta\|_1 = \sum_{j=1}^d |\theta_j|$
  - Perform model selection as well as regularization
  - Non-smooth optimization and structured sparsity
  - See, e.g., Bach, Jenatton, Mairal and Obozinski (2012a,b)

"old style" Supervised learning

- Data: n observations $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, i.i.d.
- Prediction as a linear function $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathbb{R}^d$
- (regularized) empirical risk minimization: find $\hat{\theta}$ solution of

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} \ell\big(Y_i, \langle \theta, \Phi(X_i) \rangle\big) \text{ such that } \Omega(\theta) \leq D$$

convex data fitting term + constraint

"old style" Supervised learning

- Data: n observations $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \ldots, n$, i.i.d.
- Prediction as a linear function $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathbb{R}^d$
- (regularized) empirical risk minimization: find $\hat{\theta}$ solution of

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} \ell\big(Y_i, \langle \theta, \Phi(X_i) \rangle\big) \text{ such that } \Omega(\theta) \leq D$$

convex data fitting term + constraint

- Empirical risk: $\hat{f}(\theta) = n^{-1} \sum_{i=1}^{n} \ell(Y_i, \langle \theta, \Phi(X_i) \rangle)$

"old style" Supervised learning

- Data: n observations $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, i.i.d.
- Prediction as a linear function $\langle \theta, \Phi(x) \rangle$ of features $\Phi(x) \in \mathbb{R}^d$
- (regularized) empirical risk minimization: find $\hat{\theta}$ solution of

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} \ell\big(Y_i, \langle \theta, \Phi(X_i) \rangle\big) \text{ such that } \Omega(\theta) \leq D$$

convex data fitting term + constraint

- Empirical risk: $\hat{f}(\theta) = n^{-1} \sum_{i=1}^{n} \ell(Y_i, \langle \theta, \Phi(X_i) \rangle)$
- Expected risk: $f(\theta) = \mathbb{E}[\ell(Y, \langle \theta, \Phi(X) \rangle)]$.

General assumptions

- Data: n observations $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, i.i.d.
- Bounded features $\Phi(x) \in \mathbb{R}^d$: $\|\Phi(x)\|_2 \leq R$
- Empirical risk $\hat{f}(\theta) = n^{-1} \sum_{i=1}^{n} \ell(Y_i, \langle \theta, \Phi(X_i) \rangle)$
- Expected risk $f(\theta) = \mathbb{E}[\ell(Y, \langle \theta, \Phi(X) \rangle)]$
- Loss for a single observation: $f_i(\theta) = \ell(Y_i, \langle \theta, \Phi(X_i) \rangle)$. For all i, $f(\theta) = \mathbb{E}[f_i(\theta)]$
- Properties of $f_i, f, \hat{f}$
  - Convex on $\mathbb{R}^d$
  - Additional regularity assumptions: Lipschitz-continuity, smoothness and strong convexity

Lipschitz continuity

- Bounded gradients of g ($\Leftrightarrow$ Lipschitz-continuity): the function g if convex, differentiable and has gradients uniformly bounded by B on the ball of center 0 and radius D: for all $\theta \in \mathbb{R}^d$,

$$\|\theta\|_2 \leq D \Rightarrow \|\nabla g(\theta)\|_2 \leq B$$
$$\Leftrightarrow$$
$$|g(\theta) - g(\theta')| \leq B\|\theta - \theta'\|_2$$

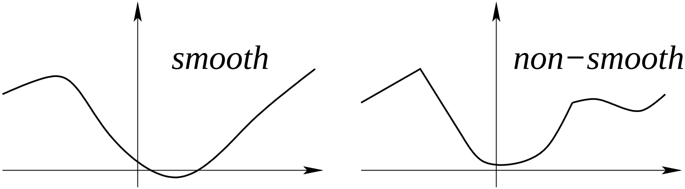
- Machine learning
  - $g(\theta) = n^{-1} \sum_{i=1}^{n} \ell(Y_i, \langle \theta, \Phi(X_i) \rangle)$
  - G-Lipschitz loss and R-bounded data: $B = GR$

Smoothness

- A function $g : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if and only if it is differentiable and its gradient is L-Lipschitz: for all $\theta, \theta' \in \mathbb{R}^d$;

$$\|\nabla g(\theta_1) - \nabla g(\theta')\|_2 \leq L\|\theta - \theta'\|_2$$

- If g is twice differentiable, for all $\theta \in \mathbb{R}^d$, $\nabla^{\otimes 2} g(\theta) \preccurlyeq L \cdot \mathrm{Id}$



Smoothness

- A function $g : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if and only if it is differentiable and its gradient is L-Lipschitz: for all $\theta, \theta' \in \mathbb{R}^d$;

$$\|\nabla g(\theta_1) - \nabla g(\theta')\|_2 \leq L\|\theta - \theta'\|_2$$

- If g is twice differentiable, for all $\theta \in \mathbb{R}^d$, $\nabla^{\otimes 2} g(\theta) \preccurlyeq L \cdot \mathrm{Id}$

Machine learning

- $g(\theta) = n^{-1} \sum_{i=1}^n \ell(Y_i, \langle \theta, \Phi(X_i) \rangle)$
- Hessian $\approx$ covariance matrix

$$n^{-1} \sum_{i=1}^n \Phi(X_i) \Phi^\top(X_i) \ddot{\ell}(Y_i, \langle \theta, \Phi(X_i) \rangle)$$

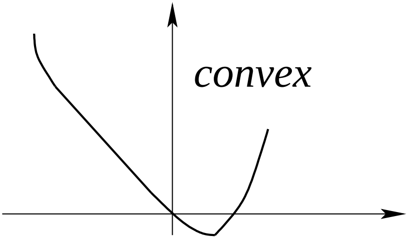
- L_{loss}-smooth loss and R-bounded data: $L = L_{\mathrm{loss}} R^2$

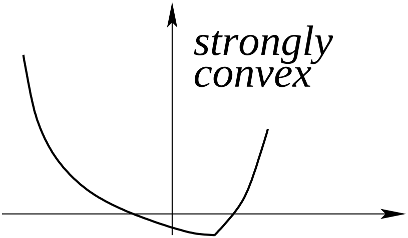
Strong convexity

- A function $g : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex if and only if, for all $\theta, \theta' \in \mathbb{R}^d$,

$$g(\theta) \geqslant g(\theta') + \langle \nabla g(\theta'), \theta - \theta' \rangle + \frac{\mu}{2} \|\theta - \theta'\|_2^2$$

- If g is twice differentiable: for all $\theta \in \mathbb{R}^d$, $\nabla^2 g(\theta) \succcurlyeq \mu \cdot \mathrm{Id}$





Strong convexity

- A function $g : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex if and only if, for all $\theta, \theta' \in \mathbb{R}^d$,

$$g(\theta) \geqslant g(\theta') + \langle \nabla g(\theta'), \theta - \theta' \rangle + \frac{\mu}{2} \|\theta - \theta'\|_2^2$$

- If g is twice differentiable: for all $\theta \in \mathbb{R}^d$, $\nabla^2 g(\theta) \succcurlyeq \mu \cdot \mathrm{Id}$

Machine learning

- $g(\theta) = n^{-1} \sum_{i=1}^n \ell(Y_i, \langle \theta, \Phi(X_i) \rangle)$
- Hessian $\approx$ covariance matrix

$$n^{-1} \sum_{i=1}^n \Phi(X_i) \Phi(X_i)^\top \ddot{\ell}(Y_i, \langle \theta, \Phi(X_i) \rangle)$$

- Data with invertible covariance matrix

Strong convexity

- A function $g : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex if and only if, for all $\theta, \theta' \in \mathbb{R}^d$,

$$g(\theta) \geqslant g(\theta') + \langle \nabla g(\theta'), \theta - \theta' \rangle + \frac{\mu}{2} \|\theta - \theta'\|_2^2$$

- If g is twice differentiable: for all $\theta \in \mathbb{R}^d$, $\nabla^2 g(\theta) \succcurlyeq \mu \cdot \mathrm{Id}$

Machine learning

- $g(\theta) = n^{-1} \sum_{i=1}^n \ell(Y_i, \langle \theta, \Phi(X_i) \rangle)$
- Hessian $\approx$ covariance matrix

$$n^{-1} \sum_{i=1}^n \Phi(X_i) \Phi(X_i)^\top \ddot{\ell}(Y_i, \langle \theta, \Phi(X_i) \rangle)$$

- Data with invertible covariance matrix

Adding regularization by $\frac{\mu}{2}\|\theta\|^2$ [! creates a bias (controlled by μ)]

Smoothness/convexity assumptions: summary

- Bounded gradients of g (Lipschitz-continuity): the function g if convex, differentiable and has gradients uniformly bounded by B on the ball of center 0 and radius D:

$$\text{for all } \theta \in \mathbb{R}^d, \|\theta\|_2 \leq D \Rightarrow \|\nabla g(\theta)\|_2 \leq B$$

- Smoothness of g: the function g is convex, differentiable with L-Lipschitz-continuous gradient ∇g:

$$\text{for all } \theta, \theta' \in \mathbb{R}^d, \quad \|\nabla g(\theta) - \nabla g(\theta')\|_2 \leq L\|\theta - \theta'\|_2$$

- Strong convexity of g: The function g is strongly convex with respect to the norm $\|\cdot\|_2$, with convexity constant $\mu > 0$: for all $\theta, \theta'` \in \mathbb{R}^d$,

$$g(\theta) \geqslant g(\theta') + \langle \nabla g(\theta'), \theta - \theta' \rangle + \frac{\mu}{2}\|\theta - \theta'\|_2^2$$

Empirical risk minimization: rationale

- The expected risk $f(\theta) = \mathbb{E}[\ell(Y, \langle\theta, X, \rangle\rangle]$ is not tractable.
- Only the empirical risk $\hat{f}(\theta) = n^{-1}\sum_{i=1}^{n}[\ell(Y_i, \langle\theta, X_i, \rangle\rangle]$ is.
- Minimizing $\hat{f}$ instead of f?
- A simple observation:

$$f(\hat{\theta}) - \min_{\theta\in\Theta} f(\theta) \leq \sup_{\theta\in\Theta}\{\hat{f}(\theta) - f(\theta)\} + \sup_{\theta\in\Theta}\{f(\theta) - \hat{f}(\theta)\}$$

- Can we have a bound on $\sup_{\theta\in\Theta}|\hat{f}(\theta) - f(\theta)|$?

Motivation from least-squares

- For least-squares, we have $\ell(y, \langle\theta, \Phi(x)\rangle) = \frac{1}{2}(y - \langle\theta, \Phi(x)\rangle)^2$, and

$$f(\theta) - \hat{f}(\theta) = \frac{1}{2}\theta^\top \left(\frac{1}{n}\sum_{i=1}^n \Phi(X_i)\Phi(X_i)^\top - \mathbb{E}\Phi(X)\Phi(X)^\top \right)\theta$$

$$- \theta^\top \left(\frac{1}{n}\sum_{i=1}^n Y_i\Phi(X_i) - \mathbb{E}Y\Phi(X) \right) + \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^n Y_i^2 - \mathbb{E}Y^2 \right)$$

$$\sup_{\|\theta\|_2 \leq D} |f(\theta) - \hat{f}(\theta)| \leq \frac{D^2}{2}\left\| \frac{1}{n}\sum_{i=1}^n \Phi(X_i)\Phi(X_i)^\top - \mathbb{E}\Phi(X)\Phi(X)^\top \right\|_{\mathrm{op}}$$

$$+ D\left\| \frac{1}{n}\sum_{i=1}^n Y_i\Phi(X_i) - \mathbb{E}Y\Phi(X) \right\|_2 + \frac{1}{2}\left| \frac{1}{n}\sum_{i=1}^n Y_i^2 - \mathbb{E}Y^2 \right|,$$

$$\sup_{\|\theta\|_2 \leq D} |f(\theta) - \hat{f}(\theta)| \leq O(1/\sqrt{n}) \text{ with high probability}$$

Slow rate for supervised learning

Assumptions (f is the expected risk, $\hat{f}$ the empirical risk)

- $\Omega(\theta) = \|\theta\|_2$ (Euclidean norm)
- "Linear" predictors: $\phi_\theta(x) = \langle \theta, \Phi(x) \rangle$, with $\|\Phi(x)\|_2 \leq R$
- G-Lipschitz loss: $f(\theta) = \ell(Y, \langle \theta, \Phi(X) \rangle)$ is GR-Lipschitz on $\Theta = \{\|\theta\|_2 \leq D\}$
- No convexity assumption

Slow rate for supervised learning

Assumptions (f is the expected risk, $\hat{f}$ the empirical risk)

- $\Omega(\theta) = \|\theta\|_2$ (Euclidean norm)
- "Linear" predictors: $\phi_\theta(x) = \langle \theta, \Phi(x) \rangle$, with $\|\Phi(x)\|_2 \leq R$
- G-Lipschitz loss: $f(\theta) = \ell(Y, \langle \theta, \Phi(X) \rangle)$ is GR-Lipschitz on $\Theta = \{\|\theta\|_2 \leq D\}$
- No convexity assumption

High-probability bounds: With probability greater than $1 - \delta$,

$$\sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \leq \frac{\sup |\ell(Y,0)| + GRD}{\sqrt{n}} \left[2 + \sqrt{2 \log \frac{2}{\delta}} \right]$$

Slow rate for supervised learning

Assumptions (f is the expected risk, $\hat{f}$ the empirical risk)

- $\Omega(\theta) = \|\theta\|_2$ (Euclidean norm)
- "Linear" predictors: $\phi_\theta(x) = \langle \theta, \Phi(x) \rangle$, with $\|\Phi(x)\|_2 \leq R$
- G-Lipschitz loss: $f(\theta) = \ell(Y, \langle \theta, \Phi(X) \rangle)$ is GR-Lipschitz on $\Theta = \{\|\theta\|_2 \leq D\}$
- No convexity assumption

Risk bounds

$$\mathbb{E}\Big[\sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)|\Big] \leq \frac{4 \sup |\ell(Y, 0)| + 4GRD}{\sqrt{n}}$$

Slow rate for supervised learning

Assumptions (f is the expected risk, $\hat{f}$ the empirical risk)

- $\Omega(\theta) = \|\theta\|_2$ (Euclidean norm)
- "Linear" predictors: $\phi_\theta(x) = \langle \theta, \Phi(x) \rangle$, with $\|\Phi(x)\|_2 \leq R$
- G-Lipschitz loss: $f(\theta) = \ell(Y, \langle \theta, \Phi(X) \rangle)$ is GR-Lipschitz on $\Theta = \{\|\theta\|_2 \leq D\}$
- No convexity assumption

Method

- Tools: Symmetrization, Rademacher complexity (see Boucheron et al., 2012), McDiarmid inequality.
- Lipschitz functions $\Rightarrow$ slow rate

Empirical Risk vs Fluctuation

- We have, with probability $1-\delta$, for all $\theta \in \Theta$:

$$f(\hat{\theta}) - \min_{\theta \in \Theta} f(\theta) \leq \sup_{\theta \in \Theta}\{\hat{f}(\theta) - f(\theta)\} + \sup_{\theta \in \Theta}\{f(\theta) - \hat{f}(\theta)\}$$
$$\leq \frac{2}{\sqrt{n}}(\ell_0 + GRD)(4 + \sqrt{2 \log \frac{1}{\delta}})$$

- Only need to optimize with precision $\approx 1/\sqrt{n}$

Slow rate for supervised learning

Assumptions (f is the expected risk, $\hat{f}$ the empirical risk)

- $\Omega(\theta) = \|\theta\|_2$ (Euclidean norm)
- "Linear" predictors: $\phi_\theta(x) = \langle \theta, \Phi(x) \rangle$, with $\|\Phi(x)\|_2 \leq R$ a.s.
- G-Lipschitz loss: f and $\hat{f}$ are GR-Lipschitz on $\Theta = \{\|\theta\|_2 \leq D\}$
- No assumptions regarding convexity
- With probability greater than $1 - \delta$

$$\sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \leq \frac{\ell_0 + GRD}{\sqrt{n}} \left[2 + \sqrt{2 \log \frac{2}{\delta}} \right]$$

- Expected estimation error: $\mathbb{E}\big[\sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \big] \leq \frac{4\ell_0 + 4GRD}{\sqrt{n}}$
- Under other conditions on the model, can we improve the rate $1/\sqrt{n}$?

Motivation from mean estimation

Estimator

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Z_i = \arg\min_{\theta \in \mathbb{R}} \hat{f}(\theta)$$

where

$$\hat{f}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(Z_i - \theta)^2 \quad f(\theta) = \mathbb{E}\Big[(Z-\theta)^2\Big]$$

Slow rate

$$f(\theta) = \frac{1}{2}(\theta - \mathbb{E}[Z])^2 + \frac{1}{2}\,\mathrm{var}(Z) = \hat{f}(\theta) + O(n^{-1/2})$$

Motivation from mean estimation

Estimator

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Z_i = \arg\min_{\theta \in \mathbb{R}} \hat{f}(\theta)$$

where

$$\hat{f}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(Z_i - \theta)^2 \quad f(\theta) = \mathbb{E}\Big[(Z-\theta)^2\Big]$$

Fast rate

$$f(\hat{\theta}) - f(\mathbb{E}[Z]) = \frac{1}{2}(\hat{\theta} - \mathbb{E}[Z])^2$$

$$\mathbb{E}\big[f(\hat{\theta}) - f(\mathbb{E}[Z])\big] = \frac{1}{2}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z]\right)^2 = \frac{1}{2n}\operatorname{var}(Z)$$

Bound only at $\hat{\theta}$ + strong convexity

Fast rate for supervised learning

Assumptions (f is the expected risk, $\hat{f}$ the empirical risk)

- Same as before (bounded features, Lipschitz loss) + strong convexity

For any $a > 0$, with probability greater than $1 - \delta$, for all $\theta \in \mathbb{R}^d$,

$$f(\hat{\theta}) - \min_{\eta \in \mathbb{R}^d} f(\eta) \leq \frac{8(1 + a^{-1})G^2 R^2 (32 + \log(\delta^{-1}))}{\mu n}$$

- Results from (Sridharan et al., 2008), (Boucheron et al., 2012).
- Strongly convex functions $\Rightarrow$ fast rate

Minimization of the expected and empirical risk

- Conclusion: $\boxed{\hat{\theta} \in \arg\min_{\theta \in \Theta} \hat{f}(\theta)}$ is a good proxy as a minimizer of f as n is large.
- Question: How to find $\hat{\theta}$?
- Answer: gradient descent algorithms!
- Recall $\hat{f}$ is assumed to be convex.
- Very efficient methods from convex optimization are available.

Complexity results in convex optimisation

- Assumption: g convex on $\mathbb{R}^d$
- Classical generic algorithms
  - (sub)gradient method/descent
  - Accelerated gradient descent
  - Newton method
- Key additional properties of g
  - Lipschitz continuity, smoothness or strong convexity
- Key insight from (Bottou and Bousquet, 2008)
  - In machine learning, no need to optimize below estimation error
- Key references: (Nesterov, 2004), (Bubeck, 2015).

(smooth) gradient descent - strong convexity

- Assumptions
  - g convex with L-Lipschitz gradient
  - g μ-strongly convex
- Algorithm:

$$\theta_t = \theta_{t-1} - \frac{1}{L} \nabla g(\theta_{t-1})$$

- Bound:

$$g(\theta_t) - g(\theta_*) \leq (1 - \mu/L)^t \big\{ g(\theta_0) - g(\theta_*) \big\}$$

(smooth) gradient descent

- Assumptions
  - g convex with L-Lipschitz gradient
  - Minimum attained at θ_*
- Algorithm:

$$\theta_t = \theta_{t-1} - \frac{1}{L} \nabla g(\theta_{t-1})$$

- Bound:

$$g(\theta_t) - g(\theta_*) \leq \frac{2L\|\theta_0 - \theta_*\|^2}{t+4}$$

- Not best possible convergence rate

Key properties of smooth convex functions

$g : \mathbb{R}^d \to \mathbb{R}$ a convex L-smooth function: for all $\theta, \eta \in \mathbb{R}^d$,

$$\|\nabla g(\theta) - \nabla g(\eta)\| \leq L\|\theta - \eta\|$$

- Quadratic upper bound

$$0 \leq g(\theta) - g(\eta) - \langle \nabla g(\eta), \theta - \eta \rangle \leq (L/2)\|\theta - \eta\|^2$$

- Co-coercivity

$$\frac{1}{L}\|\nabla g(\theta) - \nabla g(\eta)\|^2 \leq \langle \nabla g(\theta) - \nabla g(\eta), \theta - \eta \rangle$$

Co-coercivity: proof

$$\frac{1}{L}\|\nabla g(\theta) - \nabla g(\eta)\|^2 \leq \langle \nabla g(\theta) - \nabla g(\eta), \theta - \eta \rangle$$

Set $\eta \in \mathbb{R}^d$ and consider the auxiliary function

$$\theta \mapsto h(\theta) = g(\theta) - \langle \nabla g(\eta), \theta \rangle \quad \nabla h(\theta) = \nabla g(\theta) - \nabla g(\eta)$$

Convex, global minimum at η and L-smooth.
Using the quadratic upper bound for h at θ, we get for any $\theta \in \mathbb{R}^d$,

$$h(\eta) \leq h\left(\theta - \frac{1}{L}\nabla h(\theta)\right) \leq h(\theta) - \frac{1}{L}\|\nabla h(\theta)\|^2 + \frac{1}{2L}\|\nabla h(\theta)\|^2$$
$$\leq h(\theta) - \frac{1}{2L}\|\nabla h(\theta)\|^2$$

Co-coercivity: proof

$$\frac{1}{L}\|\nabla g(\theta) - \nabla g(\eta)\|^2 \leq \langle \nabla g(\theta) - \nabla g(\eta), \theta - \eta \rangle$$

Partial conclusion:

$$h(\eta) \leq h(\theta) - \frac{1}{2L}\|\nabla h(\theta)\|^2$$

with $h(\theta) = g(\theta) - \langle \nabla g(\eta), \theta \rangle$.

$$g(\eta) - \langle \nabla g(\eta), \eta \rangle \leq g(\theta) - \langle \nabla g(\eta), \theta \rangle - \frac{1}{2L}\|\nabla g(\theta) - \nabla g(\eta)\|^2$$

Co-coercivity: proof

$$\frac{1}{L}\|\nabla g(\theta) - \nabla g(\eta)\|^2 \leq \langle \nabla g(\theta) - \nabla g(\eta), \theta - \eta \rangle$$

$$g(\eta) - \langle \nabla g(\eta), \eta \rangle \leq g(\theta) - \langle \nabla g(\eta), \theta \rangle - \frac{1}{2L}\|\nabla g(\theta) - \nabla g(\eta)\|^2$$

Conclusion:

$$g(\theta) \geq g(\eta) + \langle \nabla g(\eta), \theta - \eta \rangle + \frac{1}{2L}\|\nabla g(\theta) - \nabla g(\eta)\|^2 .$$

Co-coercivity: proof

Adding

$$g(\theta) \geq g(\eta) + \langle \nabla g(\eta), \theta - \eta \rangle + \frac{1}{2L} \|\nabla g(\theta) - \nabla g(\eta)\|^2$$
$$g(\eta) \geq g(\theta) + \langle \nabla g(\theta), \eta - \theta \rangle + \frac{1}{2L} \|\nabla g(\theta) - \nabla g(\eta)\|^2$$

we obtain

$$\frac{1}{L} \|\nabla g(\theta) - \nabla g(\eta)\|^2 \leq \langle \nabla g(\theta) - \nabla g(\eta), \theta - \eta \rangle$$

Smooth Strongly convex functions

g is L-smooth and μ-strongly convex.

- Two key properties:
  - Strong convexity: $\langle \nabla g(\theta) - \nabla g(\eta), \theta - \eta \rangle \geq \mu \|\theta - \eta\|^2$
  - Smoothness: $\|\nabla g(\theta) - \nabla g(\eta)\| \geq L \|\theta - \eta\|$
- The value $Q_g = L/\mu$ is the condition number of g.

Smooth Strongly convex functions

- Strong convexity optimality certificate

$$g(\theta) \leq g(\eta) + \langle \nabla g(\eta), (\theta - \eta) \rangle + \frac{1}{2\mu} \|\nabla g(\theta) - \nabla g(\eta)\|^2 .$$

- Strong co-coercivity

$$\langle \nabla g(\theta) - \nabla g(\eta), \theta - \eta \rangle \geq \frac{\mu L}{\mu + L} \|\theta - \eta\|^2 + \frac{1}{\mu + L} \|\nabla g(\theta) - \nabla g(\eta)\|^2 .$$

Proof of the upper bound for strongly convex functions

$$g(\theta) \leq g(\eta) + \langle \nabla g(\eta), \theta - \eta \rangle + \frac{1}{2\mu} \left\| \nabla g(\theta) - \nabla g(\eta) \right\|^2$$

- $h : \theta \mapsto h(\theta) = g(\theta) - \langle \nabla g(\eta), \theta \rangle$ is strongly convex with a global minimum at η.
- Since h is strongly convex, for all $\theta, \zeta \in \mathbb{R}^d$, we get

$$h(\zeta) \geq h(\theta) + \langle \nabla h(\theta), \zeta - \theta \rangle + \frac{\mu}{2} \|\zeta - \theta\|^2 .$$

- Hence, for all $\theta \in \mathbb{R}^d$,

$$\begin{aligned} h(\eta) = \min_{\zeta} h(\zeta) &\geq \min_{\zeta} \left\{ h(\theta) + \langle \nabla h(\theta), \zeta - \theta \rangle + \frac{\mu}{2} \|\zeta - \theta\|^2 \right\} \\ &\geq h(\theta) - \frac{1}{2\mu} \|\nabla h(\theta)\|^2 \end{aligned}$$

Proof of the upper bound for strongly convex functions

$$g(\theta) \leq g(\eta) + \langle \nabla g(\eta), \theta - \eta \rangle + \frac{1}{2\mu} \left\| \nabla g(\theta) - \nabla g(\eta) \right\|^2$$

Optimality certificate: taking $\eta = \theta_*$ and using that

$$\nabla g(\theta_*) = 0$$

we get that for all $\theta \in \mathbb{R}^d$,

$$g(\theta) - g(\theta_*) \leq \frac{1}{2\mu} \left\| \nabla g(\theta) \right\|^2$$

Proof of strong co-coercivity

Set $h(\theta) = g(\theta) - (\mu/2)\|\theta\|^2$. We get

$$\begin{aligned}\langle \nabla h(\theta) - \nabla h(\eta), \theta - \eta \rangle &= \langle \nabla g(\theta) - \nabla g(\eta), \theta - \eta \rangle - \mu\|\theta - \eta\|^2 \\ &\leq (L - \mu)\|\theta - \eta\|^2\end{aligned}$$

Hence, h is $L - \mu$-smooth. The co-coercivity implies

$$\langle h(\theta) - h(\eta), \theta - \eta \rangle \geq \frac{1}{L - \mu}\|\theta - \eta\|^2 \, .$$

which yields the result.

Convergence proof - strongly convex functions

Iteration $\theta_t = \theta_{t-1} - \gamma \nabla g(\theta_{t-1})$ with $\gamma = 1/L$.
Quadratic Upper Bound:

$$\begin{aligned} g(\theta_t) &= g\left(\theta_{t-1} - \gamma \nabla g(\theta_{t-1})\right) \\ &\leq g(\theta_{t-1}) + \langle \nabla g(\theta_{t-1}), -\gamma \nabla g(\theta_{t-1}) \rangle + \frac{L}{2} \| - \gamma \nabla g(\theta_{t-1}) \|^2 \\ &= g(\theta_{t-1}) - \gamma(1 - \gamma L/2) \| \nabla g(\theta_{t-1}) \|^2 = g(\theta_{t-1}) - \frac{1}{2L} \| \nabla g(\theta_{t-1}) \|^2 \end{aligned}$$

Strong convexity optimality certificate

$$g(\theta_t) \leq g(\theta_{t-1}) - \frac{\mu}{L} \{ g(\theta_{t-1}) - g(\theta_*) \}$$

Convergence proof - strongly convex functions

Iteration $\theta_t = \theta_{t-1} - \gamma \nabla g(\theta_{t-1})$ with $\gamma = 1/L$.

$$g(\theta_t) \leq g(\theta_{t-1}) - \frac{\mu}{L}\{g(\theta_{t-1}) - g(\theta_*)\}$$

$$g(\theta_t) - g(\theta_*) \leq (1 - \mu/L)\{g(\theta_{t-1}) - g(\theta_*)\}$$

which implies that

$$g(\theta_t) - g(\theta_*) \leq (1 - \mu/L)^t\{g(\theta_0) - g(\theta_*)\}$$

Strongly convex functions: parameter convergence

g L-smooth and μ-strongly convex. Set $r_t^2 = \|\theta_t - \theta^*\|$. We get

$$\begin{aligned} r_{t+1}^2 &= \left\|\theta_t - \theta^* - \gamma \nabla g\left(\theta_t\right)\right\|^2 \\ &= r_t^2 - 2\gamma \left\langle \nabla g\left(\theta_t\right), \theta_t - \theta^* \right\rangle + \gamma^2 \left\|\nabla g\left(\theta_t\right)\right\|^2 \\ &\leq \left(1 - \frac{2\gamma\mu L}{\mu + L}\right) r_t^2 + \gamma \left(\gamma - \frac{2}{\mu + L}\right) \left\|\nabla g\left(\theta_t\right)\right\|^2 \end{aligned}$$

Taking $0 < \gamma \leq \frac{2}{\mu+L}$, we finally get

$$r_{t+1}^2 \leq \left(1 - \frac{2\gamma\mu L}{\mu + L}\right) r_t^2$$

Strongly convex functions: parameter convergence

If $\gamma = \frac{2}{\mu+L}$, then

$$\|\theta_t - \theta^*\| \leq \left(\frac{Q_g - 1}{Q_g + 1}\right)^t \|\theta_0 - \theta^*\|$$

$$g(\theta_t) - g^* \leq \frac{L}{2}\left(\frac{Q_g - 1}{Q_g + 1}\right)^{2t} \|\theta_0 - \theta^*\|^2$$

Convergence proof - gradient descent smooth convex function

Iteration $\theta_t = \theta_{t-1} - \gamma \nabla g(\theta_{t-1})$ with $\gamma = 1/L$.

Property: The distance to the optimum θ_* decreases !

$$\begin{aligned} \|\theta_t - \theta_*\|^2 &= \|\theta_{t-1} - \theta_* - \gamma \nabla g(\theta_{t-1})\|^2 \\ &= \|\theta_{t-1} - \theta_*\|^2 + \gamma^2 \|\nabla g(\theta_{t-1})\|^2 - 2\gamma \langle \theta_{t-1} - \theta_*, \nabla g(\theta_{t-1}) \rangle \end{aligned}$$

The co-coercivity property implies that

$$\langle \theta_{t-1} - \theta_*, \nabla g(\theta_{t-1}) \rangle \geq (1/L) \|\nabla g(\theta_{t-1})\|^2$$

showing that

$$\begin{aligned} \|\theta_t - \theta_*\|^2 &\leq \|\theta_{t-1} - \theta_*|^2 - \gamma(2/L - \gamma) \|\nabla g(\theta_{t-1})\|^2 \leq \|\theta_{t-1} - \theta_*\|^2 \\ &\leq \|\theta_0 - \theta_*\|^2 \end{aligned}$$

Convergence proof - gradient descent smooth convex function

Iteration $\theta_t = \theta_{t-1} - \gamma \nabla g(\theta_{t-1})$ with $\gamma = 1/L$.

- Quadratic upper bound:

$$g(\theta_t) \le g(\theta_{t-1}) - \frac{1}{2L} \|\nabla g(\theta_{t-1})\|^2$$

- Convexity:

$$g(\theta_{t-1}) - g(\theta_*) \le \langle \nabla g(\theta_{t-1}), \theta_{t-1} - \theta_* \rangle \le \|\nabla g(\theta_{t-1})\| \, \|\theta_{t-1} - \theta_*\|$$

- Using that $\|\theta_t - \theta_*\| \le \|\theta_0 - \theta_*\|$,

$$g(\theta_t) - g(\theta_*) \le g(\theta_{t-1}) - g(\theta_*) - \frac{1}{2L\|\theta_0 - \theta_*\|^2} \{g(\theta_{t-1}) - g(\theta_*)\}^2$$

Convergence proof - gradient descent smooth convex function

Setting

$$\Delta_t = g(\theta_t) - g(\theta_*) \quad \text{and} \quad \alpha = \frac{1}{2L\|\theta_0 - \theta_*\|^2}$$

we have to analyze the convergence of

$$\Delta_t \leq \Delta_{t-1} - \alpha \Delta_{t-1}^2$$

Quadratic upper-bound:

$$\Delta_t = g(\theta_t) - g(\theta_*) \leq (L/2)\|\theta_t - \theta_*\|^2 .$$

Convergence proof - gradient descent smooth convex function

Setting

$$\Delta_t = g(\theta_t) - g(\theta_*) \quad \text{and} \quad \alpha = \frac{1}{2L\|\theta_0 - \theta_*\|^2}$$

we have to analyze the convergence of

$$\Delta_t \leq \Delta_{t-1} - \alpha \Delta_{t-1}^2$$

$$\frac{1}{\Delta_{s-1}} \leq \frac{1}{\Delta_s} - \alpha \frac{\Delta_{s-1}}{\Delta_s} \quad \text{divide by } \Delta_s \Delta_{s-1}$$

$$\frac{1}{\Delta_{s-1}} \leq \frac{1}{\Delta_s} - \alpha \quad \text{because } \Delta_s \text{ is decreasing}$$

Convergence proof - gradient descent smooth convex function

Setting

$$\Delta_t = g(\theta_t) - g(\theta_*) \quad \text{and} \quad \alpha = \frac{1}{2L\|\theta_0 - \theta_*\|^2}$$

we have to analyze the convergence of

$$\Delta_t \leq \Delta_{t-1} - \alpha \Delta_{t-1}^2$$

$$\frac{1}{\Delta_0} \leq \frac{1}{\Delta_t} - \alpha t \quad \text{by summing for } s = 1 \text{ to } t$$

$$\Delta_t \leq \frac{\Delta_0}{1 + \alpha t \Delta_0} .$$

Using that $\alpha = \{2L\|\theta_0 - \theta_*\|^2\}^{-1}$ and $\Delta_0 \leq (L/2)\|\theta_0 - \theta_*\|^2$, yields

$$\Delta_t \leq \frac{2L\|\theta_0 - \theta_*\|^2}{t+4}$$

Limits on convergence rate of first-order methods

- First-order method: any iterative algorithm that selects θ_t in $\theta_0 + \mathrm{span}(\nabla g(\theta_0), \dots, \nabla g(\theta_{t-1}))$
- Problem class: convex L-smooth functions with a global minimizer θ_*

Theorem

For every integer $t \leq (n-1)/2$ and every $\theta_0 \in \mathbb{R}^n$ there exist a function g in the problem class such that for any first-order method, we have that

$$g(\theta_t) - g(\theta_*) \geq \frac{3L\|\theta_0 - \theta_*\|^2}{32(t+1)^2}$$

where θ_* is the minimum of the function g.

$O(1/t)$ rate for gradient method might not be optimal!

Proof of the lower bound

Consider the "worst function in the world" [Nesterov, 2004]. Set $n \in \mathbb{N}$ and for any $k \in \{1, \ldots, n\}$, consider the function

$$g_k(\theta) = \frac{L}{8}[(\theta^1)^2 + \sum_{i=1}^{k-1}(\theta^i - \theta^{i+1})^2 + (\theta^k)^2 - 2\theta^1]$$

Fact 1: g_k is convex and L-smooth:

$$\langle \nabla^2 g_k(\theta)s, s\rangle = \frac{L}{4}\left[(s^1)^2 + \sum_{i=1}^{k-1}(s^i - s^{i+1})^2 + (s^k)^2\right]$$

and

$$\langle \nabla^2 g_k(\theta)s, s\rangle \leq \frac{L}{2}\left[(s^1)^2 + 2\sum_{i=1}^{k-1}2((s^i)^2 + (s^{i+1})^2) + (s^k)^2\right] \leq L\sum_{i=1}^{k}(s^i)^2$$

Proof of the lower bound

Consider the "worst function in the world" [Nesterov, 2004]. Set $n \in \mathbb{N}$ and for any $k \in \{1, \dots, n\}$, consider the function

$$g_k(\theta) = \frac{L}{8}[(\theta^1)^2 + \sum_{i=1}^{k-1}(\theta^i - \theta^{i+1})^2 + (\theta^k)^2 - 2\theta^1]$$

Fact 2 minimizer supported by first k coordinates (closed form)

$$\bar{\theta}_k^{(i)} = \begin{cases} 1 - \frac{i}{k+1}, & i = 1, \dots, k, \\ 0, & k+1 \leq i \leq n. \end{cases}$$

and the optimal value of function g_k is

$$g_k^* = \frac{L}{8}\left(-1 + \frac{1}{k+1}\right).$$

Proof of the lower bound

Consider the "worst function in the world" [Nesterov, 2004]. Set $n \in \mathbb{N}$ and for any $k \in \{1, \ldots, n\}$, consider the function

$$g_k(\theta) = \frac{L}{8}[(\theta^1)^2 + \sum_{i=1}^{k-1}(\theta^i - \theta^{i+1})^2 + (\theta^k)^2 - 2\theta^1]$$

Fact 2 Note also that

$$\sum_{i=1}^{k} i^2 = \frac{k(k+1)(2k+1)}{6} \leq \frac{(k+1)^3}{3}$$

Therefore

$$\|\bar{\theta}_k\|^2 = \sum_{i=1}^{n}(\bar{\theta}_k^{(i)})^2 = \sum_{i=1}^{k}(1 - \frac{i}{k+1})^2$$

$$= k - \frac{2}{k+1}\sum_{i=1}^{k} i + \frac{1}{(k+1)^2}\sum_{i=1}^{k} i^2 = \frac{1}{3}(k+1) \ .$$

Proof of the lower bound

Consider the "worst function in the world" [Nesterov, 2004]. Set $n \in \mathbb{N}$ and for any $k \in \{1, \ldots, n\}$, consider the function

$$g_k(\theta) = \frac{L}{8}[(\theta^1)^2 + \sum_{i=1}^{k-1}(\theta^i - \theta^{i+1})^2 + (\theta^k)^2 - 2\theta^1]$$

Fact 3 any first-order method starting from zero will be supported in the first k coordinates after iteration k
Denote $R^{k,n} = \{\theta \in R^n | \theta^{(i)} = 0,\ k+1 \leq i \leq n\}$; that is a subspace of R^n, in which only the first k components of the point can differ from zero. From the analytical form of the functions $\{g_k\}$ it is easy to see that for all $\theta \in R^{k,n}$ we have

$$g_p(\theta) = g_k(\theta)\ ,\ p = k \ldots n.$$

Proof of the lower bound

Consider the "worst function in the world" [Nesterov, 2004]. Set $n \in \mathbb{N}$ and for any $k \in \{1, \dots, n\}$, consider the function

$$g_k(\theta) = \frac{L}{8}[(\theta^1)^2 + \sum_{i=1}^{k-1}(\theta^i - \theta^{i+1})^2 + (\theta^k)^2 - 2\theta^1]$$

Fact 3 any first-order method starting from zero will be supported in the first k coordinates after iteration k
Let us fix some $p, 1 \leq p \leq n$. Let $\theta_0 = 0$. Then for any sequence $\{\theta_k\}_{k=0}^{p}$ satisfying the condition

$$\theta_k \in \mathcal{L}_k = \mathrm{Lin}\{\nabla g_p(\theta_0), \dots, \nabla g_p(\theta_{k-1})\}$$

we have $\mathcal{L}_k \subseteq R^{k,n}$.

Proof of the lower bound

Consider the "worst function in the world" [Nesterov, 2004]. Set $n \in \mathbb{N}$ and for any $k \in \{1, \dots, n\}$, consider the function

$$g_k(\theta) = \frac{L}{8}[(\theta^1)^2 + \sum_{i=1}^{k-1}(\theta^i - \theta^{i+1})^2 + (\theta^k)^2 - 2\theta^1]$$

Fact 4 For any sequence $\{\theta_k\}_{k=0}^p$ such that $\theta_0 = 0$ and $\theta_k \in \mathcal{L}_k$ we have

$$g_p(\theta_k) \geq g_k^*.$$

Indeed, $\theta_k \in \mathcal{L}_k \subseteq R^{k,n}$ and therefore

$$g_p(\theta_k) = g_k(\theta_k) \geq g_k^*.$$

Proof of the lower bound

Consider the "worst function in the world" [Nesterov, 2004]. Set $n \in \mathbb{N}$ and for any $k \in \{1, \dots, n\}$, consider the function

$$g_k(\theta) = \frac{L}{8}[(\theta^1)^2 + \sum_{i=1}^{k-1}(\theta^i - \theta^{i+1})^2 + (\theta^k)^2 - 2\theta^1]$$

At iteration k, take $g = g_{2k+1}$ and compute a lower-bound for

$$\frac{g(\theta_k) - g(\theta_*)}{\|\theta_0 - \theta_*\|^2}$$

Accelerated gradient methods (Nesterov, 1983)

Assumptions: g convex, L-smooth $1\,\mathrm{min}$. attained at θ_*

Algorithm

$$\begin{aligned}\theta_t &= \eta_{t-1} - \frac{1}{L}\nabla g(\eta_{t-1}) \\ \eta_t &= \theta_t + \frac{t-1}{t+2}(\theta_t - \theta_{t-1})\end{aligned}$$

Bound

$$g(\theta_t) - g(\theta_*) \leq \frac{2L\|\theta_0 - \theta_*\|^2}{(t+1)^2}$$

Extension to strongly-convex functions

Assumptions: g convex, L-smooth, strongly convex
Algorithm

$$\theta_t = \eta_{t-1} - \frac{1}{L}\nabla g(\eta_{t-1})$$

$$\eta_t = \theta_t + \frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}(\theta_t - \theta_{t-1})$$

Bound

$$g(\theta_t) - g(\theta_*) \leq L\|\theta_0 - \theta_*\|^2(1-\sqrt{\mu/L})^t$$

Related to conjugate gradient for quadratic functions

Subgradient

Definition

The subgradient $\partial f(\theta)$ of f at θ is the set of vectors $\vartheta \in \mathbb{R}^d$ satisfying

$$f(\vartheta) \geq f(\theta) + \langle s, \vartheta - \theta \rangle \quad \theta, \vartheta \in \mathbb{R}^d$$

- the definition is unilateral ! the affine function $\vartheta \to f(\theta) + \langle s, \vartheta - \theta \rangle$ minorizes f and coincides with f at $\theta = \vartheta$
- The definition is global in the sense that it involves all $\vartheta \in \mathbb{R}^d$
- Seems to deviate from the "classical" concept of differentials (no remainder terms, the condition is local and not global)

Basic subgradient calculus

(a) Scaling: $\partial(af) = a\partial f$ provided $a > 0$. The condition $a > 0$ makes the function f remain convex

(b) Addition: $\partial\left(f+g\right) = \partial\left(f\right) + \partial\left(g\right)$ if int dom $f \cap \mathrm{dom}\, g \neq \emptyset$.

(c) Affine composition: if $g(\theta) = f(A\theta + b)$ then $\partial g(\theta) = A^T \partial f(A\theta + b)$.

(d) If f is differentiable at a point $\theta \in \mathrm{int}\,\mathrm{dom}\, f$ then $\partial f\left(\theta\right) = \{\nabla f\left(\theta\right)\}$.

Basic optimality conditions for convex optimization: unconstrained case

Theorem

Let f be convex. If θ is a local minimum of f, then θ is a global minimum of f. Furthermore, this happens if and only if $0 \in \partial f(\theta)$

Proof.

It can be easily seen that $0 \in \partial f(\theta)$ if and only if θ is a global minimum. Now assume that θ is a local minimum of f. Then for any η and λ small enough

$$f(\theta) \leq f((1-\lambda)\theta + \lambda\eta) \leq (1-\lambda)f(\theta) + \lambda f(\eta),$$

which implies that $f(\theta) \leq f(\eta)$ and thus that θ is a global minimum of f. $\square$

Basic optimality conditions for convex optimization: constrained case

Given a convex set $\Theta \subseteq \mathbb{R}^d$ and a convex function $f : \Theta \to \mathbb{R}$, we intend to

$$\min_{\theta \in \Theta} f(\theta)$$

Define the characteristic of the convex set Θ

$$I_\Theta(\theta) := \begin{cases} 0, & \theta \in \Theta \\ \infty & Otherwise \end{cases}$$

By definition of subgradients, the subdifferential of I_Θ is given by the normal cone at θ

$$\partial I_\Theta(\theta) = \{w \in \mathbb{R}^n \mid \langle w, \eta - \theta \rangle \leq 0, \forall \eta \in \Theta\}$$

Basic optimality conditions for convex optimization: constrained case

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function and Θ be a convex set. Then θ_* is an optimal solution of $\min_{\theta \in \Theta} f(\theta)$ if and only if there exists $w_* \in \partial f(\theta_*)$ such that

$$\langle w_*, \eta - \theta_* \rangle \geq 0, \quad \forall \eta \in \Theta \,.$$

Subgradient: links with directional derivatives

- Since for any $s \in \partial f(\theta)$, we have $f(\vartheta) \geq f(\theta) + \langle s, \vartheta - \theta \rangle$ for all $\vartheta \in \mathbb{R}^d$, for any $\zeta \in \mathbb{R}^d$ and $t \geq 0$ we get

$$t^{-1}\{f(\theta + t\zeta) - f(\theta)\} \geq \langle s, \zeta \rangle$$

- Taking the limit at $t \downarrow 0^+$, for all $\theta, \zeta \in \mathbb{R}^d$,

$$\langle s, \zeta \rangle \leq f'(\theta, \zeta)$$

Subgradient: links with directional derivatives

- Conversely, if for all $\zeta \in \mathbb{R}^d$, $\langle s, \zeta \rangle \leq f'(\theta, \zeta)$, then for all $t \geq 0$, the increase slope property implies

$$\langle s, \zeta \rangle \leq f'(\theta, \zeta) \leq t^{-1}\{f(\theta + t\zeta) - f(\theta)\}$$

- Taking $t = 1$ and $\zeta = \vartheta - \theta$,

$$f(\theta) + \langle s, \vartheta - \theta \rangle \leq f(\vartheta)$$

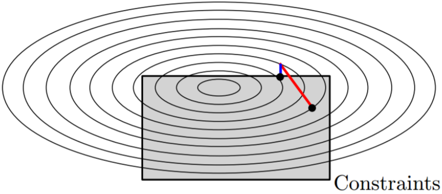
showing that $s \in \partial f(\theta)$.

$$\begin{aligned}\partial f(\theta) &= \{s \in \mathbb{R}^d : f(\theta) + \langle s, \vartheta - \theta \rangle \leq f(\vartheta) \quad \text{for all } \vartheta \in \mathbb{R}^d\} \\ &= \{s \in \mathbb{R}^d : \langle s, \zeta \rangle \leq f'(\theta, \zeta) \quad \text{for all } \zeta \in \mathbb{R}^d\}\end{aligned}$$

Subgradient method/descent

Assumptions: g convex and B-Lipschitz-continuous on $\{\|\theta\|_2 \leq D\}$

Algorithm: $\theta_t = \Pi_D\Big(\theta_{t-1} - \gamma_t \partial g(\theta_{t-1})\Big)$ where Π_D : orthogonal projection onto $\{\|\theta\|_2 \leq D\}$



Subgradient method/descent

Assumptions: g convex and B-Lipschitz-continuous on $\{\|\theta\|_2 \leq D\}$

Algorithm: $\theta_t = \Pi_D\Big(\theta_{t-1} - \gamma_t \partial g(\theta_{t-1})\Big)$ where Π_D : orthogonal projection onto $\{\|\theta\|_2 \leq D\}$

Bound [with optimally chosen stepsize γ_t]

$$g\bigg(\frac{1}{t}\sum_{k=0}^{t-1}\theta_k\bigg) - g(\theta_*) \leq \frac{2DB}{\sqrt{t}}$$

Best possible convergence rate after $O(d)$ iterations (Bubeck, 2015)

Subgradient method/descent - proof - I

Iteration: $\theta_t = \Pi_D(\theta_{t-1} - \gamma_t \partial g(\theta_{t-1}))$

Assumption: $\|\partial g(\theta)\|_2 \leq B$ and $\|\theta\|_2 \leq D$

$\|\theta_t - \theta_*\|_2^2 \leq \|\theta_{t-1} - \theta_* - \gamma_t \partial g(\theta_{t-1})\|_2^2$ by contractivity of projections

$\leq \|\theta_{t-1} - \theta_*\|_2^2 + B^2\gamma_t^2 - 2\gamma_t \langle \theta_{t-1} - \theta_*, \partial g(\theta_{t-1})\rangle$ because $\|\partial g(\theta_{t-1})\|_2 \leq B$

$\leq \|\theta_{t-1} - \theta_*\|_2^2 + B^2\gamma_t^2 - 2\gamma_t \big[g(\theta_{t-1}) - g(\theta_*)\big]$ property of subgradients

leading to

$$g(\theta_{t-1}) - g(\theta_*) \leq \frac{B^2\gamma_t}{2} + \frac{1}{2\gamma_t}\big[\|\theta_{t-1} - \theta_*\|_2^2 - \|\theta_t - \theta_*\|_2^2\big]$$

Subgradient method/descent - proof - I

$$g(\theta_{t-1}) - g(\theta_*) \leq \frac{B^2\gamma_t}{2} + \frac{1}{2\gamma_t}\left[\|\theta_{t-1} - \theta_*\|_2^2 - \|\theta_t - \theta_*\|_2^2\right]$$

Constant step-size $\gamma_t = \gamma$

$$\sum_{u=1}^{t}[g(\theta_{u-1}) - g(\theta_*)] \leq \sum_{u=1}^{t}\frac{B^2\gamma}{2} + \sum_{u=1}^{t}\frac{1}{2\gamma}\left[\|\theta_{u-1} - \theta_*\|_2^2 - \|\theta_u - \theta_*\|_2^2\right]$$

$$\leq \; t\frac{B^2\gamma}{2} + \frac{1}{2\gamma}\|\theta_0 - \theta_*\|_2^2 \leq t\frac{B^2\gamma}{2} + \frac{2}{\gamma}D^2$$

Optimal step-size $\gamma_t = \frac{2D}{B\sqrt{t}}$ depends on the horizon

Convexity: $g\left(\frac{1}{t}\sum_{k=0}^{t-1}\theta_k\right) - g(\theta_*) \leq \frac{2DB}{\sqrt{t}}$

Sub-gradient: decreasing stepsize

$$g(\theta_{t-1}) - g(\theta_*) \leq \frac{B^2\gamma_t}{2} + \frac{1}{2\gamma_t}\big[\|\theta_{t-1}-\theta_*\|_2^2 - \|\theta_t-\theta_*\|_2^2\big]$$

$$\sum_{u=1}^{t}\big[g(\theta_{u-1}) - g(\theta_*)\big] \leq \sum_{u=1}^{t}\frac{B^2\gamma_u}{2} + \sum_{u=1}^{t}\frac{1}{2\gamma_u}\big[\|\theta_{u-1}-\theta_*\|_2^2 - \|\theta_u-\theta_*\|_2^2\big]$$

$$= \sum_{u=1}^{t}\frac{B^2\gamma_u}{2} + \sum_{u=1}^{t-1}\|\theta_u-\theta_*\|_2^2\Big(\frac{1}{2\gamma_{u+1}} - \frac{1}{2\gamma_u}\Big) + \frac{\|\theta_0-\theta_*\|_2^2}{2\gamma_1} - \frac{\|\theta_t-\theta_*\|_2^2}{2\gamma_t}$$

$$\leq \sum_{u=1}^{t}\frac{B^2\gamma_u}{2} + \sum_{u=1}^{t-1}4D^2\Big(\frac{1}{2\gamma_{u+1}} - \frac{1}{2\gamma_u}\Big) + \frac{4D^2}{2\gamma_1} = \sum_{u=1}^{t}\frac{B^2\gamma_u}{2} + \frac{4D^2}{2\gamma_t}\,.$$

Convexity: with $\gamma_u = 2D/(B\sqrt{u})$ we get

$$g\left(\frac{1}{t}\sum_{k=0}^{t-1}\theta_k\right) - f(\theta_*) \leq \frac{2DB}{\sqrt{t}}$$

Subgradient descent for machine learning

Assumptions (f is the expected risk, $\hat{f}$ the empirical risk)

- "Linear" predictors: $\theta(x) = \langle \theta, \Phi(x) \rangle$, with $\|\Phi(x)\|_2 \leq R$
- $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \langle \Phi(X_i), \theta \rangle)$
- G-Lipschitz loss: f and $\hat{f}$ are GR-Lipschitz on $\Theta = \{\|\theta\|_2 \leq D\}$

High-probability bound: with probability greater than $1 - \delta$

$$\sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \leq \frac{GRD}{\sqrt{n}} \left[2 + \sqrt{2 \log \frac{2}{\delta}} \right]$$

Optimization: after t iterations of subgradient method

$$\hat{f}(\hat{\theta}) - \min_{\theta \in \Theta} \hat{f}(\theta) \leq \frac{GRD}{\sqrt{t}}$$

$t = n$ iterations, with total running-time complexity of $O(n^2 d)$

Summary: rate of convergence

Assumption g convex
Gradient descent $\theta_t = \Pi_{\mathcal{D}}\left(\theta_{t-1} - \gamma_t \partial g(\theta_{t-1})\right)$
Problem parameters

- D diameter of the domain
- B Lipschitz-constant
- L smoothness constant
- μ strong convexity constant

	convex	strongly convex
nonsmooth	deterministic: $BD/\sqrt{t}$	deterministic: B^2/t
smooth	deterministic: LD^2/t^2	deterministic: $\exp(-t\sqrt{\mu/L})$

Going back to minimization of expected and empirical risks

- From a finite set of observations: $Z_1, \ldots, Z_n$, the empirical risk:

$$\hat{f}(\theta) = (1/n) \sum_{i=1}^{n} \ell(\theta, Z_i) \; .$$

- In the case n is moderate, we can use the algorithms considered before.
- In the case
  - n is very large (say $\geq 10^6$),
  - the data is distributed among different devices,

  these methods cannot be used anymore.
- Solution: batch learning
- This method belongs to the very rich class of stochastic approximation schemes.

1 Supervised Machine Learning

2 Smooth convex optimization

3 Non-smooth convex optimization

4 Stochastic approximation

5 Proximal methods

6 Applications

Links with batch learning

Empirical risk minimization

- Finite set of observations: $Z_1, \ldots, Z_n$
- Minimize the empirical risk: $\hat{f}(\theta) = \frac{1}{n} \sum_{k=1}^{n} \ell(\theta, Z_i)$

Batch stochastic gradient

- Let $S \subset \{1, \ldots, n\}$ be a mini-batch sampled with/without replacement in $\{1, \ldots, n\}$ with cardinal $|S| = N$.
- Define the mini-batch gradient

$$\nabla \hat{f}_S(\theta) = (1/p) \sum_{i \in S} \nabla_\theta \ell(\theta, Z_i) \ ,$$

where $p = n/N$ or $p = 1/\binom{N}{n}$.

- Then, $\nabla \hat{f}_S$ is an unbiased estimator of $\nabla \hat{f}$, i.e.

$$\mathbb{E}[\nabla \hat{f}_S(\theta) | (Z_i)_{i \in \{1, \ldots, n\}}] = \nabla \hat{f}(\theta) \ .$$

Links with batch learning

Empirical risk minimization

- Minimize the empirical risk: $\hat{f}(\theta) = \frac{1}{n}\sum_{k=1}^{n} \ell(\theta, Z_i)$

Batch stochastic gradient

- Batch stochastic optimization consists in replacing $\nabla \hat{f}(\theta_k)$ by the minibatch estimate $\nabla \hat{f}_{S_{k+1}}(\theta_k)$ in the gradient descent scheme to define the iterates $(\theta_k)_{k\in\mathbb{N}}$,

$$\theta_{k+1} = \theta_k - \gamma_{k+1} \nabla \hat{f}_{S_{k+1}}(\theta_k) \ ,$$

where (S_k) is an i.i.d. sequence of minibatches and $(\gamma_k)_{k\in\mathbb{N}^*}$ is a sequence of stepsizes.

Links with batch learning

Remarks

- $(S_k)_{k\in\mathbb{N}^*}$ uniform with/without replacement non necessary the best choice.
- $(\gamma_k)_{k\in\mathbb{N}^*}$ is either held constant or decreasing going to 0:
  - If it is constant $\gamma_k \equiv \gamma$, the scheme does not converge in general: there exists a small bias of order γ;
  - If $\lim_{k\to+\infty} \gamma_k = 0$, then the scheme converge under appropriate conditions.
- This scheme belongs to the class of stochastic approximation schemes.

Links with online learning

Expected risk minimization

- Minimize the expected risk: $f(\theta) = \mathbb{E}[\ell(\theta, Z)]$

Online stochastic gradient

- Let $(Z_k)_{k\in\mathbb{N}^*}$ be an i.i.d. sequence.
- Define for any $k \in \mathbb{N}^*$,

$$\nabla f_k(\theta) = \nabla_\theta \ell(\theta, Z_k) \ .$$

- Then, ∇f_k is an unbiased estimator of ∇f, i.e.

$$\mathbb{E}[\nabla \hat{f}_k(\theta)] = \nabla \hat{f}(\theta)$$

where the expectation is taken over the data $(Z_k)_{k\in\mathbb{N}^*}$.

Links with online learning

Empirical risk minimization

- Minimize the expected risk: $f(\theta) = \mathbb{E}[\ell(\theta, Z)]$

Online stochastic gradient

- Online stochastic gradient defines the iterates $(\theta_k)_{k\in\mathbb{N}}$,

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \nabla f_{n+1}(\theta_n) \ ,$$

where $(\gamma_k)_{k\in\mathbb{N}^*}$ is a sequence of stepsizes.

Remarks

- $(\gamma_k)_{k\in\mathbb{N}^*}$ is either constant or decrease to 0.
- This scheme also belongs to the class of stochastic approximation/optimization schemes.

Stochastic gradient descent

Goal of stochastic gradient:

- Minimize a function f defined on $\mathbb{R}^d$
- given only unbiased estimates ∇f_n of ∇f,
- or ∂f_n of its subgradients ∂f.

Online learning

- loss for a single pair of observations: $\boxed{f_n(\theta) = \ell(Y_n, \langle \theta, \Phi(X_n) \rangle)}$
- $f(\theta) = \mathbb{E}[f_n(\theta)] = \mathbb{E}[\ell(Y_n, \langle \theta, \Phi(X_n) \rangle)]$ = generalization error
- Expected gradient:

$$\nabla f(\theta) = \mathbb{E}[\nabla f_n(\theta)] = \mathbb{E}[\dot{\ell}(Y_n, \langle \theta, \Phi(X_n) \rangle) \, \Phi(X_n)]$$

- Non-asymptotic results

Number of iterations = number of observations

Convex stochastic approximation

Key properties of f and/or f_n

- Smoothness: f B-Lipschitz continuous, ∇f L-Lipschitz continuous
- Strong convexity: f μ-strongly convex

Key algorithm: Stochastic (sub)gradient descent (a.k.a. Robbins-Monro)

$$\boxed{\theta_n = \theta_{n-1} - \gamma_n \nabla f_n(\theta_{n-1}) \,, \quad \theta_n = \theta_{n-1} - \gamma_n \partial f_n(\theta_{n-1})}$$

Stochastic approximation beyond convex optimization

- Stochastic approximation goes far beyond convex optimization.
- Problem: find the roots of the mean field function h, i.e. solve $h(\theta) = 0$.
- In stochastic optimization: $h = \nabla f$.
- The function h is not known in closed form, but

$$h(\theta) = \int H(\theta, x)\nu(\mathrm{d}x)$$

where $H : \Theta \times \mathsf{X} \to \Theta$ is a known function and ν is a probability distribution over X.

Stochastic approximation beyond convex optimization: Robbins Monro set up

- Assume that there is an i.i.d. sequence $\{X_n,\ n \in \mathbb{N}\}$ distributed according to ν
- The stochastic approximation procedure:

$$\theta_n = \theta_{n-1} + \gamma_n H(\theta_{n-1}, X_n) \text{ with } \mathbb{E}[h_n(\theta_{n-1})|\mathcal{F}_{n-1}] = h(\theta_{n-1})$$

where $\mathcal{F}_{n-1}$ is the σ-algebra of summarizing "past" observations.
- Can alternatively be written

$$\theta_n = \theta_{n-1} + \gamma_n h(\theta_{n-1}) + \gamma_n M_n$$

where $M_n = H(\theta_{n-1}, X_n) - h(\theta_n)$.
- Under the stated assumptions, $\mathbb{E}\left[M_n \,|\, \mathcal{F}_{n-1}\right] = 0$, i.e. the sequence $\{M_n,\ n \in \mathbb{N}\}$ is a martingale increment sequence.

Convex stochastic approximation

Key properties of f and/or f_n

- Smoothness: f B-Lipschitz continuous, ∇f L-Lipschitz continuous
- Strong convexity: f μ-strongly convex

Key algorithm: Stochastic (sub)gradient descent (a.k.a. Robbins-Monro)

$$\boxed{\theta_n = \theta_{n-1} - \gamma_n \nabla f_n(\theta_{n-1}) \ , \quad \theta_n = \theta_{n-1} - \gamma_n \partial f_n(\theta_{n-1})}$$

- Polyak-Ruppert averaging: $\bar{\theta}_n = n^{-1} \sum_{k=0}^{n-1} \theta_k$
- Which learning rate sequence γ_n? Classical setting: $\boxed{\gamma_n = C n^{-\alpha}}$

Desirable practical behavior

- Applicable (at least) to classical supervised learning problems
- Robustness to (potentially unknown) constants (L,B,μ)
- Adaptivity to difficulty of the problem (e.g., strong convexity)

Smoothness/convexity assumptions

Iteration $\theta_n = \theta_{n-1} - \gamma_n \nabla f_n(\theta_{n-1})$.

Polyak-Ruppert averaging $\overline{\theta}_n = \dfrac{1}{n} \sum_{k=0}^{n-1} \theta_k$

Strong convexity of f: The function f is strongly convex with respect to the norm $\|\|_2$ with convexity constant $\mu > 0$:

- Invertible population covariance matrix or regularization by $\dfrac{\mu}{2}\|\theta\|^2$
- there exists a unique minimizer $\theta^\star$

Smoothness of f_n: For each $n \geq 1$ the function f_n satisfies *a.s.*:

- convex;
- differentiable with L-Lipschitz-continuous gradient ∇f_n;
- bounded variance (bounded data): almost surely

$$\mathbb{E}[\|\nabla f_{n+1}(\theta^\star)\|^2 | \mathcal{F}_n] \leq \sigma^2 \ .$$

Summary of new results (Bach and Moulines, 2011-2013)

Assumptions

- Stochastic gradient descent with learning rate $\gamma_n = Cn^{-\alpha}$
- Strongly convex smooth objective functions
- Bounded variance (bounded data): w.p. 1, $\mathbb{E}[\|\nabla f_{n+1}(\theta^\star)\|^2 | \mathcal{F}_n] \leq \sigma^2$.

Results

- Old: $O(n^{-1})$ rate achieved without averaging for $\alpha = 1$
- New: $O(n^{-1})$ rate achieved with averaging for $\alpha \in [1/2,\ 1]$
- Non-asymptotic analysis with explicit constants
- Robustness to the choice of C

Summary of new results (Bach and Moulines, 2011-2013)

Assumptions

- Stochastic gradient descent with learning rate $\gamma_n = Cn^{-\alpha}$
- Strongly convex smooth objective functions
- Bounded variance (bounded data): w.p. 1, $\mathbb{E}[\|\nabla f_{n+1}(\theta^\star)\|^2|\mathcal{F}_n] \leq \sigma^2$.

Results

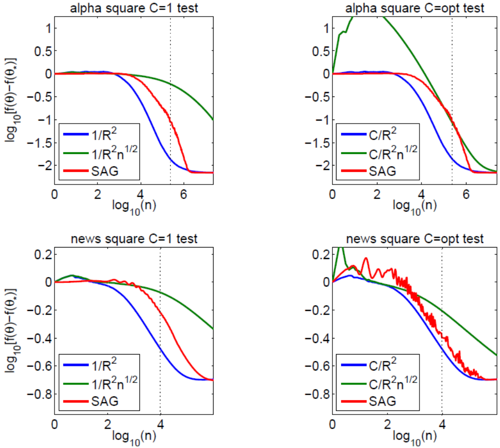
- Old: $O(n^{-1})$ rate achieved without averaging for $\alpha = 1$
- New: $O(n^{-1})$ rate achieved with averaging for $\alpha \in [1/2, \ 1]$
- Non-asymptotic analysis with explicit constants
- Robustness to the choice of C

Convergence rate for $\mathbb{E}[\|\theta_n - \theta^\star\|^2]$ and $\mathbb{E}[\|\bar{\theta}_n - \theta^\star\|^2]$.

- without averaging: $O(\gamma_n) + O(\mathrm{e}^{-\mu n \gamma_n})\|\theta_0 - \theta^\star\|^2$
- with averaging: $O(n^{-1}) + O(n^{-2\alpha}) + \mu^{-2}\|\theta_0 - \theta^\star\|^2 O(n^{-2})$

Examples

- *alpha* $(d = 500,\ n = 500\ 000)$, *news* $(d = 1\ 300\ 000,\ n = 20\ 000)$



Sketch of proof - f strongly convex, f_n smooth, bounded variance

- Consider $\delta_n = \|\theta_n - \theta^\star\|^2$.
- Then, we have almost surely

$$\delta_{n+1} = \delta_n - \gamma_{n+1}\langle \nabla f_{n+1}(\theta_n), \theta_n - \theta^\star \rangle + \gamma_{n+1}^2 \|\nabla f_{n+1}(\theta_n)\|^2 \ .$$

- f is strongly convex:

$$\begin{aligned} \mathbb{E}[\delta_{n+1}|\mathcal{F}_n] &= \delta_n - \gamma_{n+1}\langle \nabla f(\theta_n), \theta_n - \theta^\star \rangle + \gamma_{n+1}^2 \mathbb{E}[\|\nabla f_{n+1}(\theta_n)\|^2 |\mathcal{F}_n] \\ &\leq (1 - \mu\gamma_{n+1})\delta_n + \gamma_{n+1}^2 \mathbb{E}[\|\nabla f_{n+1}(\theta_n) - \nabla f(\theta^\star)\|^2 |\mathcal{F}_n] \ . \end{aligned}$$

Sketch of proof - f strongly convex, f_n smooth, bounded variance

- Consider $\delta_n = \|\theta_n - \theta^\star\|^2$.
- Then, we have almost surely

$$\delta_{n+1} = \delta_n - \gamma_{n+1}\langle \nabla f_{n+1}(\theta_n), \theta_n - \theta^\star \rangle + \gamma_{n+1}^2 \|\nabla f_{n+1}(\theta_n)\|^2 \ .$$

- Since ∇f_{n+1} is a.s. Lipschitz with bounded variance at $\theta^\star$,

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f_{n+1}(\theta_n) - \nabla f(\theta^\star)\|^2 \,\middle|\, \mathcal{F}_n \right] \\
&\leq \mathbb{E}\left[\|\nabla f_{n+1}(\theta_n) - \nabla f_{n+1}(\theta_*) + \nabla f_{n+1}(\theta_*) - \nabla f(\theta^\star)\|^2 \,\middle|\, \mathcal{F}_n \right] \\
&\leq 2(L^2\delta_n + \sigma^2) \ .
\end{aligned}$$

Sketch of proof - f strongly convex, f_n smooth, bounded variance

- Consider $\delta_n = \|\theta_n - \theta^\star\|^2$.
- Then, we have almost surely

$$\delta_{n+1} = \delta_n - \gamma_{n+1}\langle \nabla f_{n+1}(\theta_n), \theta_n - \theta^\star \rangle + \gamma_{n+1}^2 \|\nabla f_{n+1}(\theta_n)\|^2 \ .$$

- Conclusion:

$$\mathbb{E}[\delta_{n+1}|\mathcal{F}_n] \leq (1 - \mu\gamma_{n+1} + 2L^2\gamma_{n+1}^2)\delta_n + 2\sigma^2\gamma_{n+1}^2 \ .$$

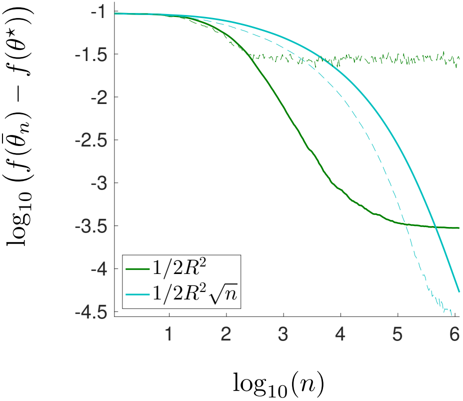
Stochastic Approximation: take home message

- Powerful algorithm:
  - Simple to implement
  - Cheap
  - No regularization needed
  - Convergence guarantees

Problems:

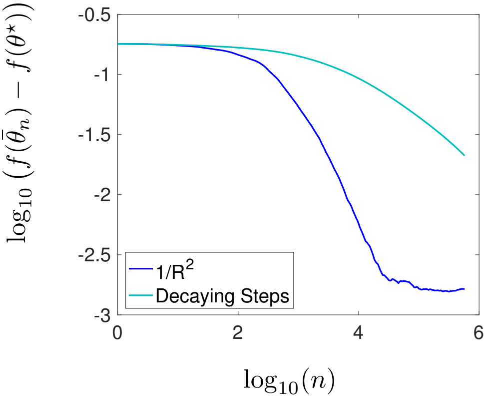
- Initial conditions can be forgotten slowly: could we use even larger/fixed step sizes?
- For fixed step sizes, the previous bounds do not show that $\mathbb{E}[\|\theta_n - \theta^\star\|^2] \not\to 0$ or $\mathbb{E}[\|\bar{\theta}_n - \theta^\star\|^2] \not\to 0$.
- We only have $\mathbb{E}[\|\theta_n - \theta^\star\|^2] = O(\gamma)$ and $\mathbb{E}[\|\bar{\theta}_n - \theta^\star\|^2] = O(\gamma)$.
- We illustrate these two facts using numerical simulations

Motivation 1/ 2. Large step sizes!



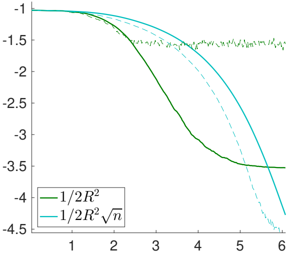
Logistic regression. Final iterate (dashed), and averaged recursion (plain).

Motivation 1/ 2. Large step sizes, real data



Logistic regression, Covertype dataset, $n = 581012$, $d = 54$. Comparison between a constant learning rate and decaying learning rate as $\frac{1}{\sqrt{n}}$.

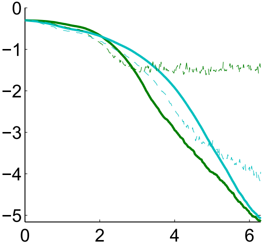
Motivation 2/ 2. Difference between quadratic and logistic loss



Logistic Regression

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta^\star) = O(\gamma^2)$$

with $\gamma = 1/(2R^2)$



Least-Squares Regression

$$\mathbb{E}f(\bar{\theta}_n) - f(\theta^\star) = O\left(\frac{1}{n}\right)$$

with $\gamma = 1/(2R^2)$

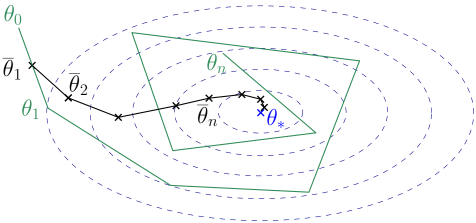
Constant learning rate SGD: convergence in the quadratic case

Least-squares: $f(\theta) = \frac{1}{2}\mathbb{E}\left[(Y - \langle \Phi(X), \theta \rangle)^2\right]$ with $\theta \in \mathbb{R}^d$

- SGD = least-mean-square algorithm
- With strong convexity assumption $\mathbb{E}\left[\Phi(X) \otimes \Phi(X)\right] = H \succcurlyeq \mu \cdot \mathrm{Id}$

$$\theta^\star = H^{-1}\mathbb{E}[Y\Phi(X)]$$

- $\bar{\theta}_n \to \theta^\star$ as $n \to +\infty$



Constant learning rate SGD: convergence in the quadratic case

- Key identity:

$$\theta_{n+1} - \theta^\star = (\mathrm{Id} - \gamma H)(\theta_n - \theta^\star) + \gamma \eta_{n+1}(\theta_n) \;, \; \mathbb{E}[\eta_{n+1}(\theta_n) | \mathcal{F}_n] = 0 \;,$$

$$\eta_{n+1}(\theta) = H\theta - \mathbb{E}[Y\Phi(X)] - \Phi(X_{n+1})\Phi(X_{n+1})^\top\theta + Y_{n+1}\Phi(X_{n+1}) \;.$$

- Therefore,

$$\theta_{n+1} - \theta^\star = (\mathrm{Id} - \gamma H)^{n+1}(\theta_0 - \theta^\star) + \gamma \sum_{k=0}^{n} (\mathrm{Id} - \gamma H)^{n-k} \eta_{k+1}(\theta_k) \;,$$

and

$$\bar{\theta}_n - \theta^\star = (n+1)^{-1} \sum_{k=0}^{n} (\theta_k - \theta^\star) \approx (n+1)^{-1} \sum_{k=0}^{n} \eta(\theta_k) \;.$$

Constant learning rate SGD: convergence in the quadratic case

Least-squares: $f(\theta) = \frac{1}{2}\mathbb{E}\left[(Y - \langle \Phi(X), \theta \rangle)^2\right]$ with $\theta \in \mathbb{R}^d$

$$\theta_{n+1} - \theta^\star = (\mathrm{Id} - \gamma H)(\theta_n - \theta^\star) + \gamma \eta_{n+1}(\theta_n) ,$$

- The sequence $(\theta_n)_{n \geq 0}$ is a homogeneous Markov chain
  1. Converges to a stationary measure π_γ
  2. $\bar{\theta}_n$ converges to $\bar{\theta}_\gamma = \int_{\mathbb{R}^d} \vartheta \mathrm{d}\pi_\gamma(\vartheta)$
- Identification of $\bar{\theta}_\gamma$
  - If $\theta_0 \sim \pi_\gamma$, then $\theta_1 \sim \pi_\gamma$.
  - Taking expectation, and using $\mathbb{E}\left[\eta_1(\theta)\right] = 0$ for any $\theta \in \mathbb{R}^d$,

$$\int_{\mathbb{R}^d} H(\vartheta - \theta^\star)\mathrm{d}\pi_\gamma(\vartheta) = 0 \Rightarrow \bar{\theta}_\gamma = \theta^\star .$$

  - Conclusion $\bar{\theta}_n \to \theta^\star$ as $n \to +\infty$ if ergodic
- Question: What happens in the general case?

SGD: an homogeneous Markov chain

- Consider a $L-$smooth and $\mu-$strongly convex function f.

- SGD with a step-size $\gamma > 0$ is an homogeneous Markov chain:

$$\theta_{k+1}^{\gamma} = \theta_k^{\gamma} - \gamma \nabla f_{k+1}(\theta_k^{\gamma}) = \theta_k^{\gamma} - \gamma \left[\nabla f(\theta_k^{\gamma}) + \eta_{k+1}(\theta_k^{\gamma}) \right] ,$$
$$\eta_{k+1}(\theta_k^{\gamma}) = \nabla f_{k+1}(\theta_k^{\gamma}) - \nabla f(\theta_k^{\gamma}) \, , \, \mathbb{E}[\eta_{k+1}(\theta_k^{\gamma}) | \mathcal{F}_n] = 0 \, .$$

Additional assumptions

- $\nabla f_k = \nabla f + \eta_{k+1}$ is almost surely L-co-coercive: for any $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\langle \nabla f_k(\theta_1) - \nabla f_k(\theta_2), \theta_1 - \theta_2 \rangle \geq L^{-1} \left\| \nabla f_k(\theta_1) - \nabla f_k(\theta_2) \right\|^2 \, .$$

- Bounded moments for p large enough,

$$\mathbb{E}[\|\epsilon_k(\theta^\star)\|^p] < \infty \, .$$

Stochastic gradient descent as a Markov Chain: Analysis framework[2]

- Let R_γ be the Markov kernel associated with $(\theta_n^\gamma)_{n\in\mathbb{N}}$.
- Existence of a stationary distribution π_γ for R_γ, and convergence to this distribution.
- Behavior under the limit distribution ($\gamma \to 0$): $\bar{\theta}_\gamma = \theta^\star + ?$
  $\looparrowright$ Provable convergence improvement with extrapolation tricks used for numerical integration and applied probability.
- Analysis of the convergence of $\bar{\theta}_n^\gamma$ to $\bar{\theta}_\gamma = \int_{\mathbb{R}^d} \vartheta \mathrm{d}\pi_\gamma(\vartheta)$ through its MSE.

[2] *Dieuleveut, D., Bach.*

Existence and convergence to a stationary distribution

Definition

Wasserstein metric: ν and λ probability measures on $\mathbb{R}^d$

$$W_2(\lambda, \nu) := \inf_{\xi \in \Pi(\lambda,\nu)} \left(\int \|\theta - \eta\|^2 \xi(\mathrm{d}\theta \cdot \mathrm{d}\eta) \right)^{1/2}$$

$\Pi(\lambda, \nu)$ is the set of probability measure ξ s.t. $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\xi(\mathsf{A} \times \mathbb{R}^d) = \lambda(\mathsf{A})$, $\xi(\mathbb{R}^d \times \mathsf{A}) = \nu(\mathsf{A})$.

Theorem

For $\gamma < L^{-1}$, the chain $(\theta_k^\gamma)_{k \geq 0}$ admits a unique stationary distribution π_γ and for all $\theta \in \mathbb{R}^d$, $n \in \mathbb{N}$:

$$W_2^2(\delta_\theta R_\gamma^n, \pi_\gamma) \leq (1 - 2\mu\gamma(1-\gamma L))^n \int_{\mathbb{R}^d} \|\theta - \vartheta\|^2 \, \mathrm{d}\pi_\gamma(\vartheta) \; .$$

Existence of a limit distribution: proof I /III

- Coupling: θ^1, θ^2 be independent and distributed according to λ_1, λ_2 respectively, and $(\theta^{(1)}_{k,\gamma})_{\geq 0}$,$(\theta^{(2)}_{k,\gamma})_{k\geq 0}$ SGD iterates:

$$\begin{cases} \theta^{(1)}_{k+1,\gamma} & = \theta^{(1)}_{k,\gamma} - \gamma\left[\nabla f(\theta^{(1)}_{k,\gamma}) + \eta_{k+1}(\theta^{(1)}_{k,\gamma})\right] \\ \theta^{(2)}_{k+1,\gamma} & = \theta^{(2)}_{k,\gamma} - \gamma\left[\nabla f(\theta^{(2)}_{k,\gamma}) + \eta_{k+1}(\theta^{(2)}_{k,\gamma})\right] \end{cases} .$$

- for all $k \geq 0$, the distribution of $(\theta^{(1)}_{k,\gamma}, \theta^{(2)}_{k,\gamma})$ is in $\Pi(\lambda_1 R^k_\gamma, \lambda_2 R^k_\gamma)$

Existence of a limit distribution: proof II/III

$$\mathbb{E}\left[\|\theta_{1,\gamma}^{(1)} - \theta_{1,\gamma}^{(2)}\|^{(2)}\right] \leq \mathbb{E}\left[\|\theta^{(1)} - \gamma\nabla f_1(\theta^{(1)}) - (\theta^{(2)} - \gamma\nabla f_1(\theta^{(2)})))\|^2\right]$$

$$\leq \mathbb{E}\left[\left\|\theta^{(1)} - \theta^{(2)}\right\|^2 - 2\gamma\left\langle\nabla f_1(\theta^{(1)}) - \nabla f_1(\theta^{(2)}), \theta^{(1)} - \theta^{(2)}\right\rangle\right]$$

$$+ \gamma^2\mathbb{E}\left[\left\|\nabla f_1(\theta^{(1)}) - \nabla f_1(\theta^{(2)})\right\|^2\right]$$

$$\overset{coco}{\leq} \mathbb{E}\left[\left\|\theta^{(1)} - \theta^{(2)}\right\|^2\right] - 2\gamma(1-\gamma L)\mathbb{E}\left[\left\langle\nabla f_1(\theta^{(1)}) - \nabla f_1(\theta^{(2)}), \theta^{(1)} - \theta^{(2)}\right\rangle\right]$$

$$\overset{unbiased}{\leq} \mathbb{E}\left[\left\|\theta^{(1)} - \theta^{(2)}\right\|^2\right] - 2\gamma(1-\gamma L)\mathbb{E}\left[\left\langle\nabla f(\theta^{(1)}) - \nabla f(\theta^{(2)}), \theta^{(1)} - \theta^{(2)}\right\rangle\right]$$

$$\overset{s.cvx.}{\leq} (1 - 2\mu\gamma(1-\gamma L))\mathbb{E}\left[\left\|\theta^{(1)} - \theta^{(2)}\right\|^2\right] .$$

Existence of a limit distribution: proof III/III

- By induction:

$$W_2^2(\lambda_1 R_\gamma^n, \lambda_2 R_\gamma^n) \leq \mathbb{E}\left[\|\theta_{n,\gamma}^{(1)} - \theta_{n,\gamma}^{(2)}\|^2\right]$$
$$\leq (1 - 2\mu\gamma(1-\gamma L))^n \int_{x,y} \|\theta_1 - \theta_2\|^2 \, \mathrm{d}\lambda_1(\theta_1)\mathrm{d}\lambda_2(\theta_2) \,.$$

- Thus $W_2(\delta_{\theta_1} R_\gamma^n, \delta_{\theta_2} R_\gamma^n) \leq (1 - 2\mu\gamma(1-\gamma L))^n \|\theta_1 - \theta_2\|^2$.
- Uniqueness, invariance, and Theorem follow:

$$W_2^2(\delta_\theta R_\gamma^n, \pi_\gamma) \leq (1 - 2\mu\gamma(1-\gamma L))^n \int_{\mathbb{R}^d} \|\theta - \vartheta\|^2 \, \mathrm{d}\pi_\gamma(\vartheta) \,.$$

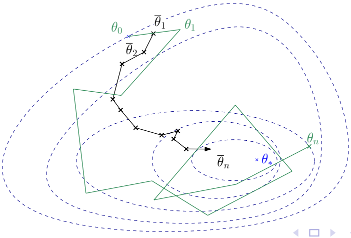
Behavior under limit distribution.

- Then we have $\mathbb{E}[\bar{\theta}_n] \to \bar{\theta}_\gamma$. Where is $\bar{\theta}_\gamma$? Close to $\theta^\star$?
- In the quadratic case $\bar{\theta}_\gamma = \theta^\star$
- In the general case, we show that

$$\bar{\theta}_\gamma = \theta^\star + \gamma \Delta(\theta^\star) + O(\gamma^2)$$

$$\Delta(\theta^\star) = f''(\theta^\star)^{-1} f'''(\theta^\star) \Big(\big[f''(\theta^\star) \otimes I + I \otimes f''(\theta^\star) \big]^{-1} \mathbb{E}[\eta(\theta^\star)^{\otimes 2}] \Big) .$$

- Linearization of the proof for the least-square

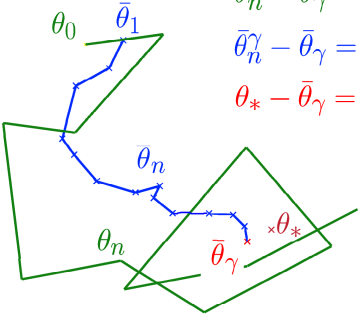


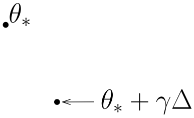
Richardson extrapolation

$$\theta_n^\gamma - \bar{\theta}_\gamma = O_p(\gamma^{1/2})$$

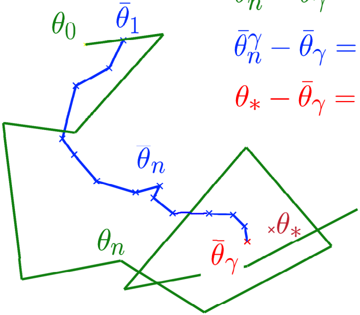
$$\bar{\theta}_n^\gamma - \bar{\theta}_\gamma = O_p(n^{-1/2})$$

$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$





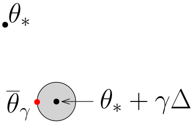
Richardson extrapolation



$$\theta_n^\gamma - \bar\theta_\gamma = O_p(\gamma^{1/2})$$

$$\bar\theta_n^\gamma - \bar\theta_\gamma = O_p(n^{-1/2})$$

$$\theta_* - \bar\theta_\gamma = O(\gamma)$$

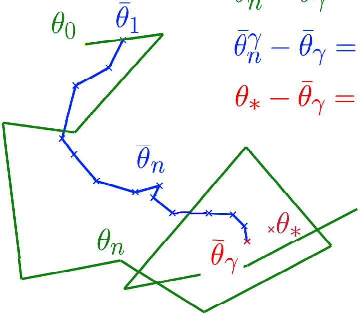


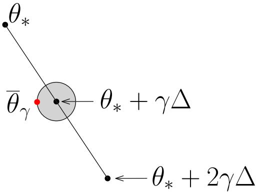
Richardson extrapolation

$$\theta_n^\gamma - \bar{\theta}_\gamma = O_p(\gamma^{1/2})$$

$$\bar{\theta}_n^\gamma - \bar{\theta}_\gamma = O_p(n^{-1/2})$$

$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$



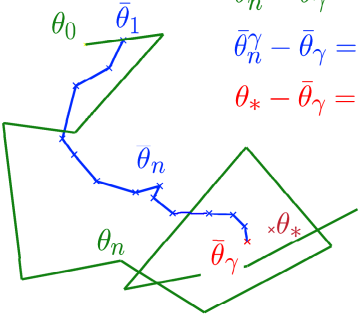


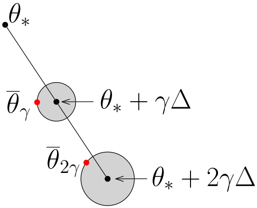
Richardson extrapolation

$$\theta_n^\gamma - \bar\theta_\gamma = O_p(\gamma^{1/2})$$

$$\bar\theta_n^\gamma - \bar\theta_\gamma = O_p(n^{-1/2})$$

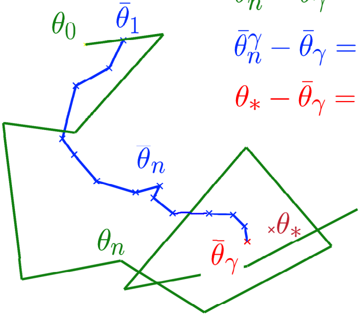
$$\theta_* - \bar\theta_\gamma = O(\gamma)$$

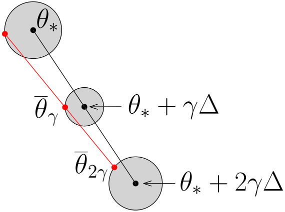




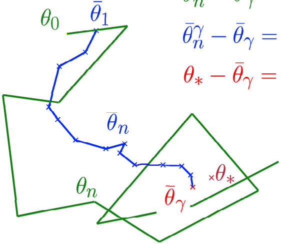
Richardson extrapolation

$$\theta_n^\gamma - \bar\theta_\gamma = O_p(\gamma^{1/2})$$
$$\bar\theta_n^\gamma - \bar\theta_\gamma = O_p(n^{-1/2})$$
$$\theta_* - \bar\theta_\gamma = O(\gamma)$$





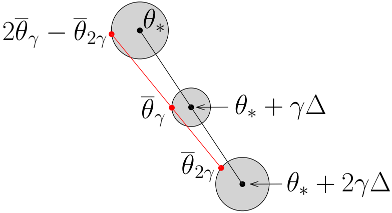
Richardson extrapolation



$$\theta_n^\gamma - \bar{\theta}_\gamma = O_p(\gamma^{1/2})$$

$$\bar{\theta}_n^\gamma - \bar{\theta}_\gamma = O_p(n^{-1/2})$$

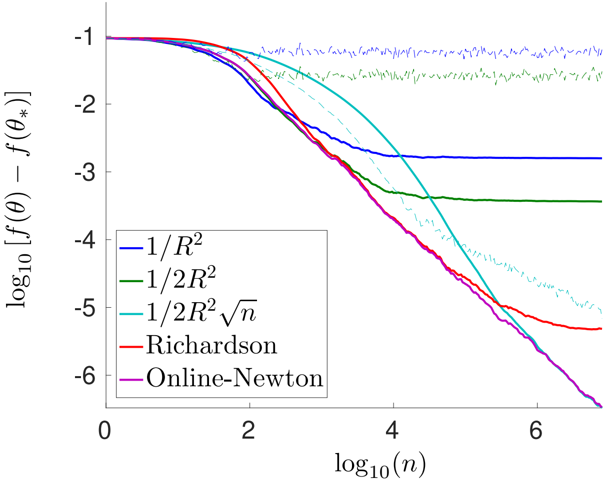
$$\theta_* - \bar{\theta}_\gamma = O(\gamma)$$



Recovering convergence closer to θ_* by **Richardson extrapolation**

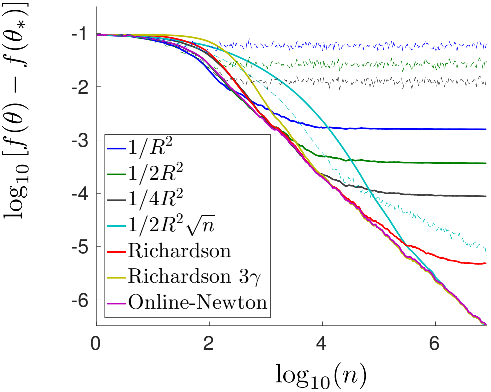
$$2\bar{\theta}_n^\gamma - \bar{\theta}_n^{2\gamma}$$

Experiments



Synthetic data, logistic regression, $n = 8\,10^6$

Experiments: Double Richardson



Synthetic data, logistic regression, $n = 8.10^6$

"Richardson 3γ": estimator built using *Richardson on 3 different sequences*: $\tilde{\theta}_n^3 = \frac{8}{3}\bar{\theta}_n^{\gamma} - 2\bar{\theta}_n^{2\gamma} + \frac{1}{3}\bar{\theta}_n^{4\gamma}$

Real data

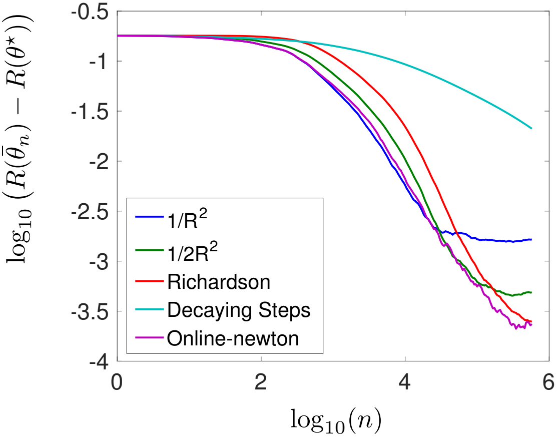


Figure: Logistic regression, Covertype dataset. $n = 581012$, $d = 54$.

Stochastic subgradient descent/method

Assumptions

- f_n convex and B-Lipschitz-continuous on $\{\|\theta\|_2 \leq D\}$
- (f_n) i.i.d. functions such that $\mathbb{E}[f_n(\theta)] = f(\theta)$
- θ_* global optimum of f on $\{\|\theta\|_2 \leq D\}$

Algorithm: $\theta_n = \Pi_D\left(\theta_{n-1} - \frac{2D}{B\sqrt{n}}\partial f_n(\theta_{n-1})\right)$

Risk Bound:

$$\mathbb{E}\left[f\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right)\right] - f(\theta_*) \leq \frac{2DB}{\sqrt{n}}.$$

- Minimax convergence rate
- Running-time complexity: $O(dn)$ after n iterations

Stochastic subgradient method - proof - I

$\theta_n = \Pi_D(\theta_{n-1} - \gamma_n \partial f_n(\theta_{n-1}))$ where $\mathcal{F}_n = \sigma((Y_k, X_k), j \leq n)$.

$\|\theta_n - \theta_*\|_2^2 \leq \|\theta_{n-1} - \theta_* - \gamma_n \partial f_n(\theta_{n-1})\|_2^2$ contractivity of projections

$\leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n \langle \theta_{n-1} - \theta_*, \partial f_n(\theta_{n-1})\rangle$ $\|\partial f_n(\theta_{n-1})\|_2 \leq B$

Taking the conditional expectations of the both sides

$$\mathbb{E}\left[\|\theta_n - \theta_*\|_2^2 | \mathcal{F}_{n-1}\right] \leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n \langle (\theta_{n-1} - \theta_*), \partial f(\theta_{n-1})\rangle$$

$\leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n \left[f(\theta_{n-1}) - f(\theta^\star)\right]$ (subgradient property)

Stochastic subgradient method - proof - I

$\theta_n = \Pi_D(\theta_{n-1} - \gamma_n \partial f_n(\theta_{n-1}))$ where $\mathcal{F}_n = \sigma((Y_k, X_k), j \leq n)$.

From

$$\mathbb{E}\left[\|\theta_n - \theta_*\|_2^2 | \mathcal{F}_{n-1}\right] \leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n \left[f(\theta_{n-1}) - f(\theta^\star)\right]$$

the tower property of conditional expectation implies

$$\mathbb{E}[\|\theta_n - \theta_*\|_2^2] \leq \mathbb{E}[\|\theta_{n-1} - \theta_*\|_2^2] + B^2\gamma_n^2 - 2\gamma_n \left[\mathbb{E}[f(\theta_{n-1})] - f(\theta^\star)\right]$$

leading to

$$\mathbb{E}[f(\theta_{n-1})] - f(\theta^\star) \leq \frac{B^2\gamma_n}{2} + \frac{1}{2\gamma_n}\left\{\mathbb{E}[\|\theta_{n-1} - \theta_*\|_2^2] - \mathbb{E}[\|\theta_n - \theta_*\|_2^2]\right\}$$

Stochastic subgradient

$$\mathbb{E}[f(\theta_{n-1})] - f(\theta^\star) \leq \frac{B^2\gamma_n}{2} + \frac{1}{2\gamma_n}\left[\mathbb{E}\|\theta_{n-1} - \theta_*\|_2^2 - \mathbb{E}\|\theta_n - \theta_*\|_2^2\right]$$

Constant step size

$$\sum_{u=1}^{n}[\mathbb{E}[f(\theta_{u-1})] - f(\theta^\star)] \leq \sum_{u=1}^{n}\frac{B^2\gamma}{2} + \sum_{u=1}^{n}\frac{1}{2\gamma}\left\{\mathbb{E}\left[\|\theta_{u-1} - \theta^\star\|_2^2\right] - \mathbb{E}\left[\|\theta_u - \theta^\star\|_2^2\right]\right\}$$
$$\leq \frac{nB^2\gamma}{2} + \frac{4D^2}{2\gamma}.$$

Optimum stepsize $\gamma = 2D/(\sqrt{n}B)$ (depends on the horizon).

Stochastic subgradient

$$\mathbb{E}[f(\theta_{n-1})] - f(\theta^\star) \leq \frac{B^2\gamma_n}{2} + \frac{1}{2\gamma_n}\left[\mathbb{E}\|\theta_{n-1} - \theta_*\|_2^2 - \mathbb{E}\|\theta_n - \theta_*\|_2^2\right]$$

Constant step size

$$\begin{aligned}\sum_{u=1}^{n}[\mathbb{E}[f(\theta_{u-1})] - f(\theta^\star)] &\leq \sum_{u=1}^{n}\frac{B^2\gamma}{2} + \sum_{u=1}^{n}\frac{1}{2\gamma}\left\{\mathbb{E}\left[\|\theta_{u-1} - \theta^\star\|_2^2\right] - \mathbb{E}\left[\|\theta_u - \theta^\star\|_2^2\right]\right\} \\ &\leq \frac{nB^2\gamma}{2} + \frac{4D^2}{2\gamma}.\end{aligned}$$

Convexity [fixed horizon]:

$$\mathbb{E}\left[f\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right)\right] - f(\theta_*) \leq \frac{2DB}{\sqrt{n}}$$

Beyond convergence in expectation

Convergence in expectation: $\mathbb{E}\left[f\left(n^{-1}\sum_{k=0}^{n-1}\theta_k\right)-f(\theta^\star)\right]\leq\frac{2DB}{\sqrt{n}}$

High-probability bounds

- Markov inequality: $\mathbb{P}\left(f\left(n^{-1}\sum_{k=0}^{n-1}\theta_k\right)-f(\theta^\star)\geq\epsilon\right)\leq\frac{2DB}{\sqrt{n}\epsilon}$
- Concentration inequality (Nemirovski et al., 2009; Nesterov and Vial, 2008)

$$\mathbb{P}\left(f\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right)-f(\theta^\star)\geq\frac{2DB}{\sqrt{n}}(2+4t)\right)\leq 2\exp(-t^2)$$

Stochastic subgradient method - proof - I

$\theta_n = \Pi_D(\theta_{n-1} - \gamma_n \partial f_n(\theta_{n-1}))$ with $\mathcal{F}_n = \sigma((Y_k, X_k), j \leq n)$.

$\|\theta_n - \theta_*\|_2^2 \leq \|\theta_{n-1} - \theta_* - \gamma_n \partial f_n(\theta_{n-1})\|_2^2$ contractivity of projections

$\leq \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n \langle \theta_{n-1} - \theta_*, \partial f_n(\theta_{n-1})\rangle$ $\|\partial f_n(\theta_{n-1})\|_2 \leq B$

Define by Z_n the error (approximation of the "true" subgradient by its noisy version)

$$Z_n = -2\langle \theta_{n-1} - \theta^\star, \partial f_n(\theta_{n-1}) - \partial f(\theta_{n-1})\rangle$$

and using the convexity we get

$$\|\theta_n - \theta^\star\|_2^2 \leq \ \|\theta_{n-1} - \theta^\star\|_2^2 + B^2\gamma_n^2 - 2\gamma_n[f(\theta_{n-1}) - f(\theta^\star)] + 2\gamma_n Z_n$$

Stochastic subgradient method - proof - II

$$Z_n = -\langle \theta_{n-1} - \theta^\star, \partial f_n(\theta_{n-1}) - \partial f(\theta_{n-1}) \rangle$$

From the inequality

$$\|\theta_n - \theta^\star\|_2^2 \leq \|\theta_{n-1} - \theta^\star\|_2^2 + B^2\gamma_n^2 - 2\gamma_n[f(\theta_{n-1}) - f(\theta^\star)] + 2\gamma_n Z_n$$

we get

$$f(\theta_{n-1}) - f(\theta^\star) \leq \frac{1}{2\gamma_n}\left\{\|\theta_{n-1} - \theta^\star\|_2^2 - \|\theta_n - \theta^\star\|_2^2\right\} + \frac{B^2\gamma_n}{2} + Z_n$$

Summing up this identity

$$\sum_{u=1}^n [f(\theta_{u-1}) - f(\theta^\star)] \leq \sum_{u=1}^n \frac{B^2\gamma_u}{2} + \sum_{u=1}^n \frac{1}{2\gamma_u}\{\|\theta_{u-1} - \theta^\star\|_2^2 - \|\theta_u - \theta^\star\|_2^2\} + \sum_{u=1}^n Z_u$$

Stochastic subgradient method - proof - II

$$Z_n = -\langle \theta_{n-1} - \theta^\star, \partial f_n(\theta_{n-1}) - \partial f(\theta_{n-1}) \rangle$$

Setting $\gamma_u = 2D/(B\sqrt{n})$ [depending on the horizon n] in

$$\sum_{u=1}^{n} [f(\theta_{u-1}) - f(\theta^\star)] \le \sum_{u=1}^{n} \frac{B^2 \gamma_u}{2} + \sum_{u=1}^{n} \frac{1}{2\gamma_u} \{\|\theta_{u-1} - \theta^\star\|_2^2 - \|\theta_u - \theta^\star\|_2^2\} + \sum_{u=1}^{n} Z_u$$

we get

$$\frac{1}{n} \sum_{u=1}^{n} \{f(\theta_{u-1}) - f(\theta^\star)\} \le \frac{2DB}{\sqrt{n}} + \frac{1}{n} \sum_{u=1}^{n} Z_u$$

Stochastic subgradient method - proof - II

$$Z_n = -\langle \theta_{n-1} - \theta^\star, \partial f_n(\theta_{n-1}) - \partial f(\theta_{n-1}) \rangle$$

Setting $\gamma_u = 2D/(B\sqrt{n})$ [depending on the horizon n] in

$$\sum_{u=1}^{n} [f(\theta_{u-1}) - f(\theta^\star)] \leq \sum_{u=1}^{n} \frac{B^2 \gamma_u}{2} + \sum_{u=1}^{n} \frac{1}{2\gamma_u} \{\|\theta_{u-1} - \theta^\star\|_2^2 - \|\theta_u - \theta^\star\|_2^2\} + \sum_{u=1}^{n} Z_u$$

we get

$$\frac{1}{n} \sum_{u=1}^{n} \{f(\theta_{u-1}) - f(\theta^\star)\} \leq \frac{2DB}{\sqrt{n}} + \frac{1}{n} \sum_{u=1}^{n} Z_u$$

Require to study $n^{-1} \sum_{k=1}^{n} Z_k$ where $(Z_k)_{k \geq 1}$ is a bounded martingale increment sequence: $|Z_k| \leq 4DB$.

Stochastic subgradient method - proof - II

$$Z_n = -\langle \theta_{n-1} - \theta^\star, \partial f_n(\theta_{n-1}) - \partial f(\theta_{n-1}) \rangle$$

Setting $\gamma_u = 2D/(B\sqrt{n})$ [depending on the horizon n] in

$$\sum_{u=1}^{n}[f(\theta_{u-1}) - f(\theta^\star)] \leq \sum_{u=1}^{n} \frac{B^2\gamma_u}{2} + \sum_{u=1}^{n} \frac{1}{2\gamma_u}\{\|\theta_{u-1} - \theta^\star\|_2^2 - \|\theta_u - \theta^\star\|_2^2\} + \sum_{u=1}^{n} Z_u$$

we get

$$\frac{1}{n}\sum_{u=1}^{n}\{f(\theta_{u-1}) - f(\theta^\star)\} \leq \frac{2DB}{\sqrt{n}} + \frac{1}{n}\sum_{u=1}^{n} Z_u$$

Azuma-Hoeffding inequality for bounded martingale increments:

$$\mathbb{P}\left(\frac{1}{n}\sum_{u=1}^{n} Z_u \geq \frac{4DBt}{\sqrt{n}}\right) \leq \exp(-t^2/2)$$

Stochastic subgradient method - proof - II

$$Z_n = -\langle \theta_{n-1} - \theta^\star, \partial f_n(\theta_{n-1}) - \partial f(\theta_{n-1}) \rangle$$

Setting $\gamma_u = 2D/(B\sqrt{n})$ [depending on the horizon n] in

$$\sum_{u=1}^{n} [f(\theta_{u-1}) - f(\theta^\star)] \leq \sum_{u=1}^{n} \frac{B^2 \gamma_u}{2} + \sum_{u=1}^{n} \frac{1}{2\gamma_u} \{\|\theta_{u-1} - \theta^\star\|_2^2 - \|\theta_u - \theta^\star\|_2^2\} + \sum_{u=1}^{n} Z_u$$

we get

$$\frac{1}{n} \sum_{u=1}^{n} \{f(\theta_{u-1}) - f(\theta^\star)\} \leq \frac{2DB}{\sqrt{n}} + \frac{1}{n} \sum_{u=1}^{n} Z_u$$

Moment bounds can be deduced from Burkholder-Rosenthal-Pinelis inequality (Pinelis, 1994)

Definition

Definition (Proximal mapping)

g: closed convex function; γ: stepsize

$$\operatorname{prox}_{\gamma,g}(\theta) = \operatorname*{argmin}_{\eta \in \Theta} \left(g(\eta) + (2\gamma)^{-1} \|\eta - \theta\|_2^2 \right)$$

- The uniqueness of the minimizer stems from the strong convexity of the function $\eta \mapsto g(\eta) + 1/(2\gamma)\|\eta - \theta\|_2^2$
- If $g = \mathbb{I}_{\mathcal{K}}$, where $\mathcal{K}$ is a closed convex set, then $\operatorname{prox}_{\gamma,g}$ is the Euclidean projection on $\mathcal{K}$

$$\operatorname{prox}_{\gamma,g}(\theta) = \operatorname*{argmin}_{\eta \in \mathcal{K}} \|\eta - \theta\|_2^2 = P_{\mathcal{K}}(\theta)$$

- The proximal operator may be seen as a generalisation of the projection on closed convex sets.

Proximal operator

Lemma

If $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ *and* $g(\theta) = \sum_{i=1}^{p} g_i(\theta_i)$, *then*

$$\mathrm{prox}_{\gamma,g}(\theta) = (\mathrm{prox}_{\gamma,g_1}(\theta_1), \mathrm{prox}_{\gamma,g_2}(\theta_2), \ldots, \mathrm{prox}_{\gamma,g_p}(\theta_p))$$

Proximal operator

Lemma

If $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ *and* $g(\theta) = \sum_{i=1}^{p} g_i(\theta_i)$, *then*

$$\mathrm{prox}_{\gamma,g}(\theta) = (\mathrm{prox}_{\gamma,g_1}(\theta_1), \mathrm{prox}_{\gamma,g_2}(\theta_2), \ldots, \mathrm{prox}_{\gamma,g_p}(\theta_p))$$

$$\begin{aligned}\operatorname*{argmin}_{(\eta_1,\ldots,\eta_p)} \left\{ \sum_{i=1}^{p} g_i(\eta_i) + 2\gamma^{-1} \sum_{i=1}^{p} \|\eta_i - \theta_i\|^2 \right\} \\ = \sum_{i=1}^{p} \operatorname*{argmin}_{\eta_i} \left\{ g_i(\eta_i) + (2\gamma)^{-1} \|\eta_i - \theta_i\|^2 \right\}\end{aligned}$$

A characterization of the proximal operator

Theorem

Let g be a convex function on Θ, $(\theta, p) \in \Theta^2$,

$$p = \mathit{prox}_{\gamma,g}(\theta) \iff \text{for all } \eta \in \Theta, \quad g(p) + \gamma^{-1}\langle \eta - p, \theta - p \rangle \leq g(\eta)$$

i.e. p is the unique element of Θ satisfying $\gamma^{-1}(\theta - p) \in \partial g(p)$.

A characterization of the proximal operator

Theorem

Let g be a convex function on Θ, $(\theta, p) \in \Theta^2$,

$$p = \mathrm{prox}_{\gamma,g}(\theta) \Longleftrightarrow \text{for all } \eta \in \Theta, \quad g(p) + \gamma^{-1}\langle \eta - p, \theta - p\rangle \leq g(\eta)$$

i.e. p is the unique element of Θ satisfying $\gamma^{-1}(\theta - p) \in \partial g(p)$.

Follows also from the characterization of the subdifferential

$$p \text{ is the minimizer of } \eta \mapsto g(\eta) + (2\gamma)^{-1}\|\eta - \theta\|_2^2$$

$$\Longleftrightarrow$$

$$0 \in \partial g(p) + \gamma^{-1}(p - \theta).$$

Proximal operator: LASSO and Elastic net

- If $g(\theta) = \sum_{i=1}^{p} \lambda_i |\theta_i|$ then $\mathrm{prox}_{\gamma,g}$ is shrinkage (soft threshold) operation

$$[S_{\lambda,\gamma}(\theta)]_i = \begin{cases} \theta_i - \gamma\lambda_i & \theta_i \geq \gamma\lambda_i \\ 0 & |\theta_i| \leq \gamma\lambda_i \\ \theta_i + \gamma\lambda_i & \theta_i \leq -\gamma\lambda_i \end{cases}$$

- If $g(\theta) = \lambda \left((1-\alpha)/2\|\theta\|_2^2 + \alpha\|\theta\|_1\right)$

$$(\mathrm{Prox}_{\gamma,g}(\tau))_i = \frac{1}{1+\gamma\lambda(1-\alpha)} \begin{cases} \tau_i - \gamma\lambda\alpha & \text{if } \tau_i \geq \gamma\lambda\alpha \\ \tau_i + \gamma\lambda\alpha & \text{if } \tau_i \leq -\gamma\lambda\alpha \\ 0 & \text{otherwise} \end{cases}$$

Fixed points of the proximal operator

Theorem

Let g be a proper convex function on Θ. The set of fixed points

$$\{\theta \in \Theta, \mathrm{prox}_{\gamma,g}(\theta) = \theta\}$$

coincide with the set of global minimum of g.

Fixed points of the proximal operator

Theorem

Let g be a proper convex function on Θ. The set of fixed points

$$\{\theta \in \Theta, \mathit{prox}_{\gamma,g}(\theta) = \theta\}$$

coincide with the set of global minimum of g.

- Characterization of the proximal point

$$\gamma^{-1}(\theta - \mathrm{prox}_{\gamma,g}(\theta)) \in \partial g(\mathrm{prox}_{\gamma,g}(\theta)).$$

- Sub-gradient: for all $\eta \in \Theta$,

$$\gamma^{-1}\langle \eta - \mathrm{prox}_{\gamma,g}(\theta), \theta - \mathrm{prox}_{\gamma,g}(\theta)\rangle + g(\mathrm{prox}_{\gamma,g}(\theta)) \leq g(\eta)$$

Conclusion

$$\theta = \mathrm{prox}_{\gamma,g}(\theta) \iff \text{for all } \eta \in \Theta, g(\mathrm{prox}_{\gamma,g}(\theta)) \leq g(\eta) .$$

Firm non-expansiveness

Theorem

If g is a proper convex function, then $\operatorname{prox}_{\gamma,g}$ and $(\mathrm{Id} - \operatorname{prox}_{\gamma,g})$ are firmly non-expansive (or co-coercive with constant 1), i.e. for all $\theta, \eta \in \Theta$,

$$\|p-q\|^2 + \|(\theta-p)-(\eta-q)\|^2 \leq \|\theta-\eta\|^2 ,$$
$$\iff \quad \langle p-q, \theta-\eta \rangle \geq \|p-q\|^2 .$$

where $p = \operatorname{prox}_{\gamma,g}(\theta)$ and $q = \operatorname{prox}_{\gamma,g}(\eta)$.

Firm non-expansiveness

Theorem

If g is a proper convex function, then $\mathrm{prox}_{\gamma,g}$ and $(\mathrm{Id} - \mathrm{prox}_{\gamma,g})$ are firmly non-expansive (or co-coercive with constant 1), i.e. for all $\theta, \eta \in \Theta$,

$$
\begin{aligned}
&\|p-q\|^2 + \|(\theta - p) - (\eta - q)\|^2 \leq \|\theta - \eta\|^2 \,, \\
\iff \quad & \langle p - q, \theta - \eta \rangle \geq \|p-q\|^2 \,.
\end{aligned}
$$

where $p = \mathrm{prox}_{\gamma,g}(\theta)$ and $q = \mathrm{prox}_{\gamma,g}(\eta)$.

$$
\gamma^{-1}\langle q - p, \theta - p\rangle + g(p) \leq g(q) \quad \gamma^{-1}\langle p - q, \eta - q\rangle + g(q) \leq g(p)
$$

Adding these two equations yield

$$
\langle p - q, (\theta - p) - (\eta - q)\rangle \geq 0 \,.
$$

Assumptions

$$\textbf{(P)} \quad \min_{\theta \in \mathbb{R}^d} F(\theta) \qquad F(\theta) = f(\theta) + g(\theta),$$

Assumptions

- $g : \mathbb{R}^d \to (-\infty, +\infty]$ closed convex
- $f : \Theta \to \mathbb{R}$ is convex continuously differentiable and ∇f is gradient Lipshitz: for all $\theta, \theta' \in \Theta$,

$$\|\nabla f(\theta) - \nabla f(\theta')\| \leq L \|\theta - \theta'\| ,$$

Proximal gradient algorithm

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1},g}\left(\theta_n - \gamma_{n+1}\nabla f(\theta_n)\right)$$

where

$$\mathrm{Prox}_{\gamma,g}(\tau) = \min_{\theta\in\Theta}\left(g(\theta) + \frac{1}{2\gamma}\|\theta-\tau\|^2\right)$$

Majorization-Minimization interpretation

- Since f is gradient Lipshitz, for all $\gamma \in (0, 1/L]$

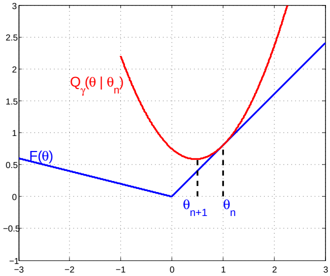
$$F(\eta) = f(\eta) + g(\eta) \leq f(\theta) + \langle \nabla f(\theta), \eta - \theta \rangle + \frac{1}{2\gamma} \|\theta - \eta\|^2 + g(\eta)$$

- Consider the following surrogate function

$$Q_\gamma(\eta|\theta) = f(\theta) + \langle \nabla f(\theta), \eta - \theta \rangle + \frac{1}{2\gamma} \|\theta - \eta\|^2 + g(\eta)$$

- For all $\theta \in \Theta$, $\eta \mapsto Q_\gamma(\eta|\theta)$ is strongly convex and has a unique minimum and

$$F(\eta) \leq Q_\gamma(\eta|\theta) \qquad\qquad F(\theta) = Q_\gamma(\theta|\theta)$$



$$F(\eta) \leq Q_\gamma(\eta|\theta_n) \qquad F(\theta_n) = Q_\gamma(\theta_n|\theta_n)$$

Majorization-Minimization interpretation

$$
\begin{aligned}
Q_\gamma(\eta|\theta) &\stackrel{\text{def}}{=} f(\theta) + \langle \nabla f(\theta), \eta - \theta \rangle + \frac{1}{2\gamma}\|\eta - \theta\|^2 + g(\eta) \\
&= f(\theta) + \frac{1}{2\gamma}\left\|\eta - (\theta - \gamma \nabla f(\theta))\right\|^2 - \frac{\gamma}{2}\|\nabla f(\theta)\|^2 + g(\eta) \,,
\end{aligned}
$$

The iterates of the proximal gradient algorithms may be rewritten as $\theta_{n+1} = T_{\gamma_{n+1}}(\theta_n)$ with the point-to-point map T_γ defined by

$$
\begin{aligned}
T_\gamma(\theta) &\stackrel{\text{def}}{=} \operatorname{Prox}_{\gamma,d}(\theta - \gamma \nabla f(\theta)) \\
&= \operatorname{argmin}_{\eta \in \operatorname{Dom}(g)} Q_\gamma(\eta|\theta) \,.
\end{aligned}
$$

Proximal gradient

- If $g(\theta) \equiv 0$, $\hookrightarrow$ gradient proximal = classical stochastic gradient

$$\theta_n = \theta_{n-1} - \gamma_n \nabla f(\theta_{n-1})$$

Proximal gradient

- If $g(\theta) \equiv 0$, $\hookrightarrow$ gradient proximal = classical stochastic gradient

$$\theta_n = \theta_{n-1} - \gamma_n \nabla f(\theta_{n-1})$$

- If $g(\theta) \equiv 0$ if $\theta \in \mathcal{C}$ and $g(\theta) = +\infty$ otherwise where $\mathcal{C}$ is a closed convex set,

$$\mathrm{Prox}_{\gamma,g}(\tau) = \min_{\theta \in \mathcal{C}} \|\tau - \theta\|^2$$

$\hookrightarrow$ gradient proximal = projected gradient

$$\theta_n = \Pi_{\mathcal{C}} \left(\theta_{n-1} - \gamma_n \nabla f(\theta_{n-1})\right)$$

Proximal gradient for the elastic net penalty

If $g(\theta) = \lambda \left(\frac{1-\alpha}{2} \|\theta\|_2^2 + \alpha \|\theta\|_1 \right)$

$$(\mathrm{Prox}_{\gamma,g}(\tau))_i = \frac{1}{1+\gamma\lambda(1-\alpha)} \begin{cases} \tau_i - \gamma\lambda\alpha & \text{if } \tau_i \geq \gamma\lambda\alpha \\ \tau_i + \gamma\lambda\alpha & \text{if } \tau_i \leq -\gamma\lambda\alpha \\ 0 & \text{otherwise} \end{cases}$$

$\hookrightarrow$ Proximal gradient= soft-thresholded gradient

$$\theta_{n+1} = \mathcal{S}_{\alpha,\lambda,\gamma_{n+1}} \left(\theta_n - \gamma_{n+1} \nabla f(\theta_n) \right)$$

Stationary points of the proximal gradient

$$\theta_{n+1} = \mathrm{Prox}_{\gamma,g}\left(\theta_n - \gamma \nabla f(\theta_n)\right) = T_\gamma(\theta_n) ,$$

where T_γ is the proximal map,

$$T_\gamma(\theta) \overset{\mathrm{def}}{=} \mathrm{Prox}_{\gamma,g}\left(\theta - \gamma \nabla f(\theta)\right) = \mathrm{argmin}_{\eta \in \mathrm{Dom}(g)} Q_\gamma(\eta|\theta) .$$

Theorem

The fixed points of the proximal map are the global minimizers of $F(\theta) = f(\theta) + g(\theta)$:

$$\mathbf{L} = \{\theta : \theta = \mathrm{Prox}_{\gamma,g}(\theta - \gamma \nabla f(\theta))\} = \{\theta \in \mathrm{Dom}(g) : 0 \in \nabla f(\theta) + \partial g(\theta)\}.$$

.

Fixed points of the proximal map

Since

$$F(\theta) = f(\theta) + g(\theta) \, ,$$

we get

$$\begin{aligned} 0 \in \partial F(\theta) \iff & \quad 0 \in \partial \gamma F(\theta) \\ \iff & \quad 0 \in \gamma \nabla f(\theta) + \partial \gamma g(\theta) \\ \iff & \quad \theta - \gamma \nabla f(\theta) \in (\theta + \gamma \partial g(\theta)) \end{aligned}$$

Fixed points of the proximal map

Since

$$F(\theta) = f(\theta) + g(\theta) \; ,$$

we get

$$\begin{aligned} 0 \in \partial F(\theta) \iff & \quad 0 \in \partial \gamma F(\theta) \\ \iff & \quad 0 \in \gamma \nabla f(\theta) + \partial \gamma g(\theta) \\ \iff & \quad \theta - \gamma \nabla f(\theta) \in (\theta + \gamma \partial g(\theta)) \end{aligned}$$

Recall that, for any η

$$p = \operatorname{prox}_{\gamma g}(\eta) \iff (\eta - p) \in \gamma \partial g(p) \iff \eta \in p + \gamma \partial g(p).$$

Fixed points of the proximal map

Since

$$F(\theta) = f(\theta) + g(\theta) \ ,$$

we get

$$\begin{aligned} 0 \in \partial F(\theta) \iff & \quad 0 \in \partial \gamma F(\theta) \\ \iff & \quad 0 \in \gamma \nabla f(\theta) + \partial \gamma g(\theta) \\ \iff & \quad \theta - \gamma \nabla f(\theta) \in (\theta + \gamma \partial g(\theta)) \end{aligned}$$

Recall that, for any η

$$p = \mathrm{prox}_{\gamma g}(\eta) \iff (\eta - p) \in \gamma \partial g(p) \iff \eta \in p + \gamma \partial g(p).$$

Hence, taking $p \leftarrow \theta$ and $\eta \leftarrow \theta - \gamma \nabla f(\theta)$

$$0 \in \partial F(\theta) \iff \theta = T_\gamma(\theta)$$

Lyapunov function

$$Q_\gamma(\eta|\theta) = f(\theta) + \langle \nabla f(\theta), \eta - \theta \rangle + \frac{1}{2\gamma}\|\theta - \eta\|^2 + g(\eta)$$

- For all $\theta \in \Theta$, $F \circ T_\gamma(\theta) \le F(\theta)$:

$$F \circ T_\gamma(\theta) \le Q_\gamma(T_\gamma(\theta)|\theta) \le Q_\gamma(\theta|\theta) = F(\theta)$$

  Moreover, the inequality is strict unless θ is a fixed point of the mapping T_γ.
- F is a Lyapunov function for the proximal map T_γ.

Convergence result

$$\textbf{(P)} \quad (\arg)\min_{\theta\in\Theta} \{f(\theta) + g(\theta)\}\,,$$

- the objective function always converge $\{F(\theta_n), n \geq 0\}$
- f is convex: then $\{\theta_n,\ n \in \mathbb{N}\}$ converges to $\theta_\star$, where $\theta_\star$ is a minimizer of F.
- $F(\theta_n) - F(\theta_\star) = O(1/n)$.
- Results similar to smooth optimization ($O(1/n)$ where n is the number of iterations)
- Acceleration methods: Nesterov, 2007; Beck and Teboulle, 2009. ($O(1/n^2)$) [algorithm FISTA]

Stochastic proximal gradient

Objective

- Exact algorithm :

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1},g}\left(\theta_n - \gamma_{n+1}\nabla f(\theta_n)\right)$$

- Pertubed algorithm :

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1},g}\left(\theta_n - \gamma_{n+1}H_{n+1}\right)$$

where H_{n+1} is a noisy approximation of the true gradient $\nabla f(\theta_n)$.

- Problem find sufficient conditions on the stochastic error

$$\eta_{n+1} = H_{n+1} - \nabla f(\theta_n)$$

to preserve convergence (closely related to SA).

Convergence of the parameter

Theorem

Assume f is L-smooth and the set $\mathbf{L} = \mathrm{argmin}_{\theta \in \Theta} F(\theta)$ is non-empty. Assume in addition that $\gamma_n \in (0, 1/L]$ for any $n \geq 1$ and $\sum_n \gamma_n = +\infty$. If the following series converge

$$\sum_{n \geq 0} \gamma_{n+1} \langle T_{\gamma_{n+1}}(\theta_n), \eta_{n+1} \rangle \;, \quad \sum_{n \geq 0} \gamma_{n+1} \eta_{n+1} \;, \quad \sum_{n \geq 0} \gamma_{n+1}^2 \|\eta_{n+1}\|^2 \;,$$

then there exists $\theta_\infty \in \mathbf{L}$ such that $\lim_n \theta_n = \theta_\infty$.

Convergence of the function

Theorem

Assume f is L-smooth and the set $\mathbf{L} = \mathrm{argmin}_{\theta \in \Theta} F(\theta)$ is non-empty. Assume that $\gamma_n \in (0, 1/L]$ and let $\{a_0, \cdots, a_n\}$ be nonnegative weights. Then, for any $\theta_\star \in \mathbf{L}$ and $n \geq 1$,

$$\sum_{k=1}^{n} a_k \left\{ F(\theta_k) - \min F \right\} \leq U_n(\theta_\star)$$

where

$$U_n(\theta_\star) \stackrel{\mathrm{def}}{=} \frac{1}{2} \sum_{k=1}^{n} \left(\frac{a_k}{\gamma_k} - \frac{a_{k-1}}{\gamma_{k-1}} \right) \|\theta_{k-1} - \theta_\star\|^2 + \frac{a_0}{2\gamma_0} \|\theta_0 - \theta_\star\|^2 \\ - \sum_{k=1}^{n} a_k \langle T_{\gamma_k}(\theta_{k-1}) - \theta_\star, \eta_k \rangle + \sum_{k=1}^{n} a_k \gamma_k \|\eta_k\|^2 \, .$$

Sanity check

- Assume that the gradient is exact, i.e. $\eta_n = 0$. Set $A_n = \sum_{k=1}^n a_k$ Then

$$F\left(A_n^{-1}\sum_{j=1}^n \theta_j\right) - \min F \leq A_n^{-1}\sum_{j=1}^n a_j F(\theta_j) - \min F$$
$$\leq \frac{1}{2}\sum_{k=1}^n \left(\frac{a_k}{\gamma_k} - \frac{a_{k-1}}{\gamma_{k-1}}\right)\|\theta_{k-1}-\theta_\star\|^2 + \frac{a_0}{2\gamma_0}\|\theta_0-\theta_\star\|^2$$

- Setting $a_k \equiv 1$ and $\gamma_k \equiv 1/L$

$$F\left(n^{-1}\sum_{j=1}^n \theta_j\right) - \min F \leq n^{-1}\sum_{j=1}^n F(\theta_j) - \min F$$
$$\leq \frac{L}{2}\|\theta_0-\theta_\star\|^2$$

- Up to constant, this is the same bound than the gradient algorithm for smooth convex function.

Perturbed gradient

- Take $a_k = \gamma_k$, for $k \in \{1, \dots, n\}$. Then, for any $\theta_\star \in \mathbf{L}$ and $n \geq 1$,

$$F\left(\Gamma_n^{-1} \sum_{k=1}^{n} \gamma_k \theta_k\right) - \min F \leq \frac{1}{2\Gamma_n} \|\theta_0 - \theta_\star\|^2 - \Gamma_n^{-1} \sum_{k=1}^{n} \gamma_k \langle T_{\gamma_k}(\theta_{k-1}) - \theta_\star, \eta_k \rangle + \Gamma_n^{-1} \sum_{k=1}^{n} \gamma_k^2 \|\eta_k\|^2 .$$

- Problem: Control the sequences $\sum_{k=1}^{n} \gamma_k \langle T_{\gamma_k}(\theta_{k-1}) - \theta_\star, \eta_k \rangle$ and $\sum_{k=1}^{n} \gamma_k^2 \|\eta_k\|^2$ in expectation or using high-probability bounds.

Robbins-Monro setting

$$\nabla f(\theta) = \int_{\mathsf{X}} H_\theta(x) \pi(\mathrm{d}x)$$

- Set
$$H_{n+1} = m_{n+1}^{-1} \sum_{j=1}^{m_{n+1}} H_{\theta_n}(X_{n+1}^{(j)})$$
where m_{n+1} is the size of the batch and $\{X_{n+1}^{(j)}, 1 \leq j \leq m_{n+1}\}$ is a sample from π independent of $\sigma(\theta_\ell, \ell \leq n)$.
- In such case,
$\mathbb{E}\left[H_{n+1} \,\middle|\, \mathcal{F}_n\right] = m_{n+1}^{-1} \sum_{j=1}^{m_{n+1}} \mathbb{E}\left[H_{\theta_n}(X_{n+1}^{(j)}) \,\middle|\, \mathcal{F}_n \right] = \nabla f(\theta_n)$ and $\eta_{n+1} = H_{n+1} - \nabla f(\theta_n)$ is a martingale increment.

Bounded case / Constant stepsizes - Risk Bounds

- Assume that $\|H_\theta(x)\| \leq B$, then $\|\eta_{n+1}\| \leq 2B$ and the stepsizes are constant $\gamma_k \equiv 1/B\sqrt{n}$ for $k \in \{1, \ldots, n\}$.
- On one hand

$$\Gamma_n^{-1} \sum_{k=1}^{n} \gamma_k^2 \|\eta_{k+1}\|^2 \leq \frac{4B}{\sqrt{n}}$$

- Risk bound: since $\mathbb{E}\left[\langle T_{\gamma_k}(\theta_{k-1}) - \theta_\star, \eta_k\rangle \,|\, \mathcal{F}_{k-1}\right] = 0$ (since $\mathbb{E}\left[\eta_k \,|\, \mathcal{F}_{k-1} = 0\right]$=0), the risk bound is

$$\mathbb{E}\left[F\left(n^{-1}\sum_{k=1}^{n}\theta_k\right)\right] - \min F \leq \frac{B}{2\sqrt{n}}\|\theta_0 - \theta_\star\|^2 + \frac{4B}{\sqrt{n}}\ .$$

- Same risk bound than the Stochastic subgradient method (minimax rate)

Bounded case / Constant stepsizes - Concentration

- Azuma-Hoeffding inequality for bounded martingale increments $\{Z_k,\ k \in \mathbb{N}^*\}$:

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n} Z_k \geq \frac{Ct}{\sqrt{n}}\right) \leq \exp(-t^2/2)$$

- Apply it to

$$Z_k = \langle T_{\gamma_k}(\theta_{k-1}) - \theta_\star, \eta_k \rangle \ .$$

Network structure estimation

- Problem fitting a discrete graphical models in a setting where the number of nodes in the graph is large compared to the sample size.
- Formalization Let A be a nonempty finite set, and $p \geq 1$ an integer. Consider a graphical model on $\mathsf{X} = \mathsf{A}^p$ with p.m.f.

$$f_\theta(x_1, \ldots, x_p) = \frac{1}{Z_\theta} \exp\left\{ \sum_{k=1}^{p} \theta_{kk} B_0(x_k) + \sum_{1 \leq j < k \leq p} \theta_{kj} B(x_k, x_j) \right\},$$

for a non-zero function $B_0 : \mathsf{A} \to \mathbb{R}$ and a symmetric non-zero function $B : \mathsf{A} \times \mathsf{A} \to \mathbb{R}$.

- The term Z_θ is the normalizing constant of the distribution (the partition function), which cannot (in general) be computed explicitly.

Network structure estimation

- Problem fitting a discrete graphical models in a setting where the number of nodes in the graph is large compared to the sample size.
- Formalization Let A be a nonempty finite set, and $p \geq 1$ an integer. Consider a graphical model on $\mathsf{X} = \mathsf{A}^p$ with p.m.f.

$$f_\theta(x_1, \ldots, x_p) = \frac{1}{Z_\theta} \exp\left\{ \sum_{k=1}^{p} \theta_{kk} B_0(x_k) + \sum_{1 \leq j < k \leq p} \theta_{kj} B(x_k, x_j) \right\},$$

for a non-zero function $B_0 : \mathsf{A} \to \mathbb{R}$ and a symmetric non-zero function $B : \mathsf{A} \times \mathsf{A} \to \mathbb{R}$.

- The real-valued symmetric matrix θ defines the graph structure and is the parameter of interest. Same interpretation as the precision matrix in a multivariate Gaussian distribution.

Network structure estimation

- Problem: Estimate θ from N realizations $\{x^{(i)}, 1 \leq i \leq N\}$ where $x^{(i)} = (x_1^{(i)}, \ldots, x_p^{(i)}) \in \mathsf{A}^p$ under sparsity constraint.
- Applications biology, social sciences,
- Main difficulty: the log-partition function $\log Z_\theta$ is intractable in general.
  - Most of the existing results use a pseudo-likelihood function.
  - One exception is [**hoefling09**], using an active set strategy (to preserve sparsity), and the junction tree algorithm for computing the partial derivatives of the log-partition function. However, this algorithm does not scale

Model

- Penalized likelihood $F(\theta) = -\ell(\theta) + g(\theta)$ where

$$\ell(\theta) = \frac{1}{N}\sum_{i=1}^{N}\langle \theta, \bar{B}(x^{(i)})\rangle - \log Z_\theta \text{ and } \quad g(\theta) = \lambda \sum_{1\leq k\leq j\leq p} |\theta_{jk}| \; ;$$

  the matrix-valued function $\bar{B} : \mathsf{X} \to \mathbb{R}^{p\times p}$ is defined by

$$\bar{B}_{kk}(x) = B_0(x_k) \qquad \bar{B}_{kj}(x) = B(x_k, x_j)\ , k \neq j\ .$$

- Intractable canonical exponential model.

Model

- Penalized likelihood $F(\theta) = -\ell(\theta) + g(\theta)$ where

$$\ell(\theta) = \frac{1}{N}\sum_{i=1}^{N}\langle \theta, \bar{B}(x^{(i)})\rangle - \log Z_\theta \text{ and } \quad g(\theta) = \lambda \sum_{1 \le k \le j \le p} |\theta_{jk}| \; ;$$

the matrix-valued function $\bar{B} : \mathsf{X} \to \mathbb{R}^{p \times p}$ is defined by

$$\bar{B}_{kk}(x) = B_0(x_k) \qquad \bar{B}_{kj}(x) = B(x_k, x_j) \; , k \neq j \; .$$

- $\theta \mapsto -\ell(\theta)$ is convex and

$$\nabla \ell(\theta) = \frac{1}{N}\sum_{i=1}^{N} \bar{B}(x^{(i)}) - \int_{\mathsf{X}} \bar{B}(z) f_\theta(z) \mu(\mathrm{d}z) \; ,$$

Implementation

- Direct simulation from the distribution f_θ is not feasible.
- If X is not too large, then a Gibbs sampler that samples from the full conditional distributions of f_θ can be easily implemented.
- Gibbs sampler is a generic algorithm that in some cases is known to mix poorly. Whenever possible we recommend the use of specialized problem-specific MCMC algorithms with better mixing properties...

Set up

- X $= \{1, \ldots, M\}$, $B_0(x) = 0$, and $B(x, y) = \mathbf{1}_{\{x=y\}}$, which corresponds to the Potts model.
- We use $M = 20$, $B_0(x) = x$, $N = 250$ and for $p \in \{50, 100, 200\}$.
- We generate the 'true' matrix θ_{true} such that it has on average p non-zero elements off-diagonal which are simulated from a uniform distribution on $(-4, -1) \cup (1, 4)$.
- All the diagonal elements are set to 0.

Algorithms

- Two versions of the stochastic proximal gradient are considered
  1. Solver 1: A version with a fixed Monte Carlo batch size $m_n = 500$, and decreasing step size $\gamma_n = \frac{25}{p} \frac{1}{n^{0.7}}$.
  2. Solver 2: A version with increasing Monte Carlo batch size $m_n = 500 + n^{1.2}$, and fixed step size $\gamma_n = \frac{25}{p} \frac{1}{\sqrt{50}}$.
- The set-up is such that both solvers draw approximately the same number of Monte Carlo samples.

Algorithms

- Two versions of the stochastic proximal gradient are considered
  1. Solver 1: A version with a fixed Monte Carlo batch size $m_n = 500$, and decreasing step size $\gamma_n = \frac{25}{p} \frac{1}{n^{0.7}}$.
  2. Solver 2: A version with increasing Monte Carlo batch size $m_n = 500 + n^{1.2}$, and fixed step size $\gamma_n = \frac{25}{p} \frac{1}{\sqrt{50}}$.
- We evaluate the convergence of each solver by computing the relative error $\|\theta_n - \theta_\infty\| / \|\theta_\infty\|$, along the iterations, where θ_∞ denotes the value returned by the solver on its last iteration.

Algorithms

- Two versions of the stochastic proximal gradient are considered
  1. Solver 1: A version with a fixed Monte Carlo batch size $m_n = 500$, and decreasing step size $\gamma_n = \frac{25}{p}\frac{1}{n^{0.7}}$.
  2. Solver 2: A version with increasing Monte Carlo batch size $m_n = 500 + n^{1.2}$, and fixed step size $\gamma_n = \frac{25}{p}\frac{1}{\sqrt{50}}$.
- We compare the optimizer output to θ_∞, not θ_{true}. Ideally, we would like to compare the iterates to the solution of the optimization problem. However in the present setting a solution is not available in closed form (and there could be more than one solution).

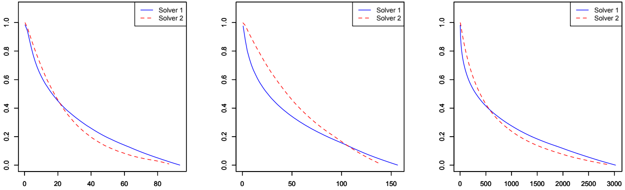


Figure: Relative errors plotted as function of computing time for Solver 1 and Solver 2.

When measured as function of resource used, Solver 1 and Solver 2 have roughly the same convergence rate.

Sensitivity and Precision

- We also compute the statistic $\mathsf{F}_n \stackrel{\text{def}}{=} \frac{2\mathsf{Sen}_n \mathsf{Prec}_n}{\mathsf{Sen}_n + \mathsf{Prec}_n}$ which measures the recovery of the sparsity structure of θ_∞ along the iteration.
- In this definition Sen_n is the sensitivity, and Prec_n is the precision defined as

$$\mathsf{Sen}_n = \frac{\sum_{j<i} \mathbf{1}_{\{|\theta_{n,ij}|>0\}} \mathbf{1}_{\{|\theta_{\infty,ij}|>0\}}}{\sum_{j<i} \mathbf{1}_{\{|\theta_{\infty,ij}|>0\}}}$$

$$\mathsf{Prec}_n = \frac{\sum_{j<i} \mathbf{1}_{\{|\theta_{n,ij}|>0\}} \mathbf{1}_{\{|\theta_{\infty,ij}|>0\}}}{\sum_{j<i} \mathbf{1}_{\{|\theta_{\infty,ij}|>0\}}}.$$

Sensitivity and Precision

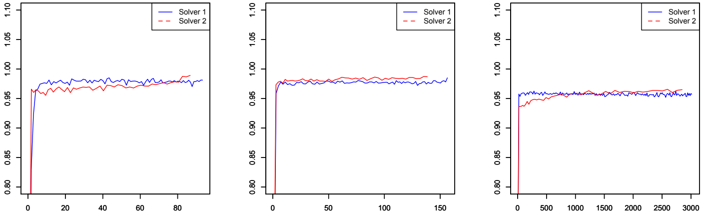


Figure: Statistic F_n plotted as function of computing time for Solver 1 and Solver 2.

High-dimensional logistic regression with random effects

- Observations : N observations $\mathbf{Y} \in \{0,1\}^N$
- Random effect : Conditionally to $\mathbf{U}$, for all $i = 1, \cdots, N$,

$$Y_i \stackrel{\text{ind.}}{\sim} \mathcal{B}\left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right)$$

where

$$\begin{bmatrix} \eta_1 \\ \vdots \\ \eta_N \end{bmatrix} = \mathbf{X}\beta_* + \sigma_* \mathbf{Z}\mathbf{U}$$

- The regressors $\mathbf{X} \in \mathbb{R}^{N \times p}$ and the factor loadings $\mathbf{Z} \in \mathbb{R}^{N \times q}$, known.
- Objective: estimate $\beta_* \in \mathbb{R}^p, \sigma_* > 0$.

Penalized likelihood

- log-likelihood : Taking $\mathbf{U} \sim \mathcal{N}_q(0, I)$, setting

$$\theta = (\beta, \sigma) \qquad F(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

the log-likelihood of the observations $\mathbf{Y}$ (with respect to θ) is

$$\ell(\theta) = \log \int \prod_{i=1}^{N} \{F(\mathbf{X}_{i \cdot}\beta + \sigma(\mathbf{ZU})_i)\}^{Y_i} \{1 - F(\mathbf{X}_{i \cdot}\beta + \sigma(\mathbf{ZU})_i)\}^{1-Y_i} \phi(\mathbf{u}) \mathrm{d}\mathbf{u}$$

- Elastic net penalty

$$g_{\lambda,\theta}(\theta) = \lambda \left(\frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right)$$

$$\tilde{g}_{\mathcal{C}}(\theta) = \begin{cases} 0 & \text{si } \theta \in \mathcal{C} \\ +\infty & \text{otherwise} \end{cases}$$

Penalized likelihood

$$\min_{\theta\in\Theta} \left(f(\theta)+g(\theta)\right) \ , \quad f(\theta) = -\ell(\theta) \ ,$$

with

$$\ell(\theta) = \log \int \exp\left(\ell_c(\theta|\mathbf{u})\right) \ \phi(\mathbf{u})\mathrm{d}\mathbf{u}$$

$$\ell_c(\theta|\mathbf{u}) = \sum_{i=1}^{N} \left\{Y_i\left(\mathbf{X}_{i\cdot}\beta + \sigma(\mathbf{Z}\mathbf{U})_i\right) - \ln\left(1+\exp\left(\mathbf{X}_{i\cdot}\beta + \sigma(\mathbf{Z}\mathbf{U})_i\right)\right)\right\}$$

Gradient :

$$\nabla\ell(\theta) = \int \nabla\ell_c(\theta|\mathbf{u})\pi_\theta(\mathbf{u})\mathrm{d}\mathbf{u}$$

where $\pi_\theta(\mathbf{u})$ is the posterior distribution of the random effect given the observations

$$\pi_\theta(\mathbf{u}) = \exp\left(\ell_c(\theta|\mathbf{u}) - \ell(\theta)\right) \ \phi(\mathbf{u})$$

Penalized likelihood

$$\min_{\theta \in \Theta} \left(f(\theta) + g(\theta) \right) \ , \quad f(\theta) = -\ell(\theta)$$

where

$$g_{\lambda,\theta}(\theta) = \lambda \left(\frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right) + \mathbb{I}_{\mathcal{C}}(\theta)$$

$$\mathbb{I}_{\mathcal{C}}(\theta) = \begin{cases} 0 & \text{if } \theta \in \mathcal{C} \\ +\infty & \text{otherwise} \end{cases} \qquad \mathcal{C} \text{ compact convex set}$$

$\hookrightarrow$ proper convex,
lower-semi continuous, not differentiable.

MCMC algorithm

- The distribution π_θ is sampled using the MCMC sampler proposed in (Polson et al, 2012) based on data-augmentation.
- We write $-\nabla \ell(\theta) = \int_{\mathbb{R}^q \times \mathbb{R}^N} H_\theta(\mathbf{u}) \tilde{\pi}_\theta(\mathbf{u}, \mathbf{w}) \, \mathrm{d}\mathbf{u} \mathrm{d}\mathbf{w}$ where $\tilde{\pi}_\theta(\mathbf{u}, \mathbf{w})$ is defined for $\mathbf{u} \in \mathbb{R}^q$ and $\mathbf{w} = (w_1, \cdots, w_N) \in \mathbb{R}^N$ by

$$\tilde{\pi}_\theta(\mathbf{u}, \mathbf{w}) = \left(\prod_{i=1}^N \bar{\pi}_{\mathsf{PG}} \left(w_i; x_i' \beta + \sigma z_i' \mathbf{u} \right) \right) \pi_\theta(\mathbf{u}) \; ;$$

- in this expression, $\bar{\pi}_{\mathsf{PG}}(\cdot; c)$ is the density of the Polya-Gamma distribution on the positive real line with parameter c given by

$$\bar{\pi}_{\mathsf{PG}}(w; c) = \cosh(c/2) \, \exp\left(-wc^2/2\right) \, \rho(w) \mathbb{1}_{\mathbb{R}+}(w) \; ,$$

where $\rho(w) \propto \sum_{k \geq 0} (-1)^k (2k+1) \exp(-(2k+1)^2/(8w)) w^{-3/2}$

MCMC algorithm

- The distribution π_θ is sampled using the MCMC sampler proposed in (Polson et al, 2012) based on data-augmentation.
- We write $-\nabla \ell(\theta) = \int_{\mathbb{R}^q \times \mathbb{R}^N} H_\theta(\mathbf{u}) \tilde{\pi}_\theta(\mathbf{u}, \mathbf{w}) \, \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{w}$ where $\tilde{\pi}_\theta(\mathbf{u}, \mathbf{w})$ is defined for $\mathbf{u} \in \mathbb{R}^q$ and $\mathbf{w} = (w_1, \cdots, w_N) \in \mathbb{R}^N$ by

$$\tilde{\pi}_\theta(\mathbf{u}, \mathbf{w}) = \left(\prod_{i=1}^N \bar{\pi}_{\mathsf{PG}} \left(w_i; x_i' \beta + \sigma z_i' \mathbf{u} \right) \right) \pi_\theta(\mathbf{u}) \; ;$$

- Thus, we have

$$\tilde{\pi}_\theta(\mathbf{u}, \mathbf{w}) = C_\theta \phi(\mathbf{u}) \prod_{i=1}^N \exp \left(\sigma (Y_i - 1/2) z_i' \mathbf{u} - w_i (x_i' \beta + \sigma z_i' \mathbf{u})^2 / 2 \right) \rho(w_i) \mathbb{1}$$

where $\ln C_\theta = -N \ln 2 - \ell(\theta) + \sum_{i=1}^N (Y_i - 1/2) x_i' \beta$.

MCMC algorithm

- The distribution π_θ is sampled using the MCMC sampler proposed in (Polson et al, 2012) based on data-augmentation.
- We write $-\nabla \ell(\theta) = \int_{\mathbb{R}^q \times \mathbb{R}^N} H_\theta(\mathbf{u}) \tilde{\pi}_\theta(\mathbf{u}, \mathbf{w}) \, \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{w}$ where $\tilde{\pi}_\theta(\mathbf{u}, \mathbf{w})$ is defined for $\mathbf{u} \in \mathbb{R}^q$ and $\mathbf{w} = (w_1, \cdots, w_N) \in \mathbb{R}^N$ by

$$\tilde{\pi}_\theta(\mathbf{u}, \mathbf{w}) = \left(\prod_{i=1}^N \bar{\pi}_{\mathsf{PG}} \left(w_i; x_i' \beta + \sigma z_i' \mathbf{u} \right) \right) \pi_\theta(\mathbf{u}) \; ;$$

- This target distribution can be sampled using a Gibbs algorithm

Numerics

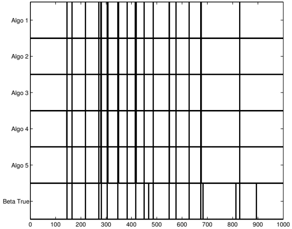
- We test the algorithms with $N = 500$, $p = 1,000$ and $q = 5$.
- We generate the $N \times p$ covariates matrix X columnwise, by sampling a stationary $\mathbb{R}^N$-valued autoregressive model with parameter $\rho = 0.8$ and Gaussian noise $\sqrt{1-\rho^2}\,\mathcal{N}_N(0, I)$.
- We generate the vector of regressors β_{true} from the uniform distribution on $[1, 5]$ and randomly set 98% of the coefficients to zero.
- The variance of the random effect is set to $\sigma^2 = 0.1$.

Numerics

We first illustrate the ability of Monte Carlo Proximal Gradient algorithms to find a minimizer of F. We compare the Monte Carlo proximal gradient algorithm

1. with fixed batch size: $\gamma_n = 0.01/\sqrt{n}$ and $m_n = 275$ (Algo 1); $\gamma_n = 0.5/n$ and $m_n = 275$ (Algo 2).
2. with increasing batch size: $\gamma_n = \gamma = 0.005$, $m_n = 200 + n$ (Algo 3); $\gamma_n = \gamma = 0.001$, $m_n = 200 + n$ (Algo 4); and $\gamma_n = 0.05/\sqrt{n}$ and $m_n = 270 + \lceil \sqrt{n} \rceil$ (Algo 5).

Results



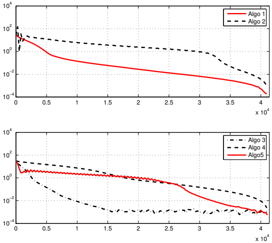


Figure: [left] The support of the sparse vector β_∞ obtained by Algo 1 to Algo 5; for comparison, the support of β_{true} is on the bottom row. [right] Relative error along one path of each algorithm as a function of the total number of Monte Carlo samples.

Results

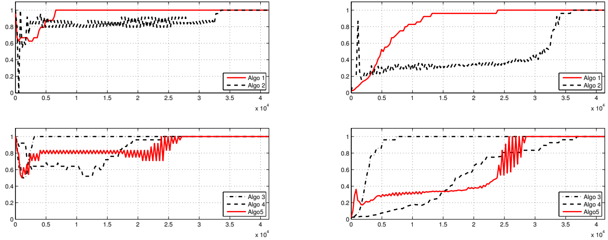


Figure: The sensitivity Sen_n [left] and the precision Prec_n [right] along a path, versus the total number of Monte Carlo samples up to time n

Bibliography I