cf. Francis Bach website

Part VII - Stochastic optimization

1 ∝ Intro

Database $S = \{(u_i, z_i) \in \mathcal{U} \times \mathcal{Z} \mid \forall i = 1, \dots n\}$

→ Assumption: prediction = linear function

$$p_x(u_i) = \langle x, \phi(u_i) \rangle$$

$$\min_x \underbrace{\frac{1}{n} \sum_{i=1}^{n} l(z_i, \langle x, \phi(u_i) \rangle) + \lambda \Omega(x)}_{F: \mathbb{R}^d \to \mathbb{R} \quad (\text{if } \phi: \mathcal{U} \to \mathbb{R}^d)}$$

→ The predictor p_x is chosen to minimize the empirical risk over a parametrized set of predictors, potentially with regularization

→ F : objective function composed with empirical risk + regularization.

→ no closed form for the minimizer of F.

2 ∝ Empirical risk versus Expected risk.

→ Empirical risk = Training cost ← the quantity you can observe

$$\frac{1}{n} \sum_{i=1}^{n} l(z_i, p_x(u_i))$$

→ Expected risk = Testing cost ← the quantity you really care about "data not seen"

$$\mathbb{E}_{(u,z)} \{ l(z, p_x(u)) \}$$

Goal : minimize the training objective but the error of unseen data

Remark : Two fundamental questions.

optimization question

① • Compute $\hat{x} \in \operatorname{Argmin} F(x)$

② • Analyze $\hat{x}$: guarantee $F(x)$ is good to minimize the expected risk.

statistical question

3 • Stochastic gradient descent (SGD)

→ Standard gradient descent requires to compute the full gradient ∇F.

Limitations : { can be costly
require to access the entire data

Batch gradient = all dataset is considered For every step of the gradient

→ Alternative : compute unbiased stochastic approximation of the gradient. $g_k(x_k)$ such that :

$$\mathbb{E}\{ g_k(x_{k-1}) \mid x_{k-1} \} = \nabla F(x_{k-1})$$

Algo :

Set $(\gamma_k)_{k \geq 0}$

Set $x_0 \in \mathbb{R}^d$

$(\forall k \geq 0) \quad x_k = x_{k-1} - \gamma_k \nabla g_k(x_{k-1})$

→ SGD for empirical risk minimization

$$F(x) = \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, p_x(u_i)) \qquad F_i(x)$$

At each iteration, choose uniformly random $i(k) = \{1, \dots n\}$ and

$$g_k(x_k) = \nabla F_{i(k)}(x_k)$$

Remark : mini-batch variant = Fixed number of training examples which is less than the dataset

* at each iteration, averaged gradient over a random subset of indices

$$x_{k+1} = x_k - \frac{\gamma_k}{n} \sum_{i \in \mathcal{I}} \nabla F_i(x_k)$$

↳ Variance reduction of the gradient estimate but use more gradient so increase running time

Remark : To avoid overfitting { regularization term / early stopping.

→ SGD for expected risk minimization

* Expected risk = Testing cost ← the quantity you really care about "data not seen"

$$\mathbb{E}_{(u,z)} \{ \ell(z, p_x(u)) \}$$

* Stochastic approximation : we assume to observe a noisy version of the gradient = random pair of observations.

expectation of unseen data and we just see samples →

$$F(x) = \mathbb{E}\{F_i(x)\} = \mathbb{E}\{\ell(z_i, p_x(u_i))\}$$

expected gradient →

$$\nabla F(x) = \mathbb{E}\{\nabla F_i(x)\}$$

(Rq : swap $\mathbb{E}\{\cdot\}$ and ∇ unbiasedness ✗ pass otherwise it create dependences.)

4. Convergence of SGD (Robbins-Monro algo.)

Ass :
- F convex and B-Lipschitz
- F admits a minimizer $\hat{x}$ s.t $\|x_0 - \hat{x}\|_2 \leq D$
- Unbiased gradient $\mathbb{E}\{g_k(x_{k-1}) \mid x_{k-1}\} = \nabla F(x_{k-1})$
- Bounded gradient $\|g_k(x_{k-1})\|_2^2 \leq B^2$ $\forall k$ almost everywhere

Set $\gamma_k = \frac{D}{B\sqrt{k}}$, SGD on F satisfies

$$\mathbb{E}\{F(\overline{x}_k) - F(x_*)\} \leq DB \frac{2+\log(k)}{\sqrt{k}}$$

where $\overline{x}_k = \left(\sum_{s=1}^{k} \gamma_s x_{s-1}\right) \Big/ \sum_{s=1}^{k} \gamma_s$.

Proof :

$$\mathbb{E}\{\|x_k - \hat{x}\|_2^2\} = \mathbb{E}\{\|x_{k-1} - \gamma_k g_k(x_{k-1}) - \hat{x}\|_2^2\}$$

$$= \mathbb{E}\{\|x_{k-1} - \hat{x}\|_2^2\} + \gamma_k^2 \mathbb{E}\{\|g_k(x_{k-1})\|^2\} - 2\gamma_k \mathbb{E}\{g_k(x_{k-1})^T (x_{k-1} - \hat{x})\}$$

$$\mathbb{E}\{g_k(x_{k-1})^T (x_{k-1} - \hat{x})\} = \mathbb{E}\{\mathbb{E}\{g_k(x_{k-1})^T (x_{k-1} - \hat{x}) \mid x_{k-1}\}\}$$

$$= \mathbb{E}\{\mathbb{E}\{g_k(x_{k-1}) \mid x_{k-1}\}^T (x_{k-1} - \hat{x})\}$$

$$= \mathbb{E}\{\nabla F(x_{k-1})^T (x_{k-1} - \hat{x})\}$$

$$\mathbb{E}\{\|x_k - \hat{x}\|_2^2\} \leq \mathbb{E}\{\|x_{k-1} - \hat{x}\|_2^2\} - 2\gamma_k \mathbb{E}\{\nabla F(x_{k-1})^T (x_{k-1} - \hat{x})\} + \gamma_k^2 B^2$$

$$\gamma_k \mathbb{E}\{\nabla F(x_{k-1})^T (x_{k-1} - \hat{x})\} \leq \frac{1}{2}\mathbb{E}\{\|x_{k-1} - \hat{x}\|_2^2\} - \frac{1}{2}\mathbb{E}\{\|x_k - \hat{x}\|^2\} + \frac{1}{2}\gamma_k^2 B^2$$

By convexity analysis $F(x_{h-1}) - F(\hat{x}) \leq \nabla F(x_{h-1})^T (x_{h-1} - \hat{x})$

- F convex $\Rightarrow$ $\cdot \leq -$

$$\alpha f(x) + (1-\alpha) f(y) \geq f(\alpha x + (1-\alpha) y)$$

$$f(x) - f(y) \geq \frac{f(\alpha x + (1-\alpha) y) - f(y)}{\alpha} \quad \xrightarrow{\lim_{\substack{\alpha \to 0 \\ \alpha \neq 0}}}$$

$$f(x) - f(y) \geq \langle \nabla f(y), x - y \rangle$$

$$f(y) - f(x) \leq \langle \nabla f(y), y - x \rangle$$

- F convex $\Leftarrow$ $\cdot \leq -$

$(x, y) \in \text{dom} F \times \text{dom} F$

$\alpha \in [0,1] \quad \alpha x + (1-\alpha) y \in \text{dom} F$

(1) $f(x) \geq f(\alpha x + (1-\alpha) y) + \langle \nabla f(\alpha x + (1-\alpha) y), \underbrace{\alpha x + (1-\alpha) y - x}_{(1-\alpha)(y-x)} \rangle$

(2) $f(y) \geq f(\alpha x + (1-\alpha) y) + \langle \nabla f(\alpha x + (1-\alpha) y, \alpha (y-x) \rangle$

(1) + (2) $\quad \alpha f(x) + (1-\alpha) f(y) \geq f(\alpha x + (1-\alpha) y)$

$$\gamma_h \, \mathbb{E}\{F(x_{h-1}) - F(\hat{x})\} \leq \frac{1}{2} \mathbb{E}\{\|x_{h-1} - \hat{x}\|^2\} - \frac{1}{2} \mathbb{E}\{\|x_h - \hat{x}\|^2\} + \frac{1}{2} \gamma_h^2 B^2$$

$$\frac{1}{\sum_{s=1}^{t} \gamma_s} \sum_{s=1}^{t} \gamma_s \, \mathbb{E}\{F(x_{s-1}) - F(\hat{x})\} \leq \frac{1}{2 \sum_{s=1}^{t} \gamma_s} \mathbb{E}\{\|x_0 - \hat{x}\|^2\} + \frac{\sum_{s=1}^{t} \gamma_s^2 B^2}{2 \sum_{s=1}^{t} \gamma_s}$$

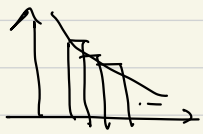
larger than $\min_{0 \leq s \leq t-1} F(x_s) - F(\hat{x})$

and than $F(\bar{x}_t) - F(\hat{x})$ with $\bar{x}_t = \frac{\sum_{s=1}^{t} \gamma_s x_{s-1}}{\sum_{s=1}^{t} \gamma_s}$ (Jensen Inequality)

goes to 0 if $\sum_{s=1}^{t} \gamma_s \to +\infty$ and $\gamma_t \to 0$

Recall : Series Integral comparisons

- $\sum_{s=1}^{t} \frac{1}{\sqrt{s}} \geq \int_0^t \frac{ds}{\sqrt{s+1}} = \left[2\sqrt{s+1}\right]_0^t = 2\sqrt{t+1} - 2 \geq \frac{1}{2}\sqrt{t}$



- $\sum_{s=1}^{t} \frac{1}{s} \leq 1 + \sum_{s=2}^{t} \frac{1}{s} \leq 1 + \int_1^t \frac{ds}{s} = 1 + \log(t)$

Choose $\gamma_s = \frac{\sigma}{\sqrt{s}}$ $\quad \sigma > 0$

$$\mathbb{E}\left\{F(\overline{x_n}) - F(\hat{x})\right\} \leq \frac{1}{2\sum_{s=1}^{n} \frac{\sigma}{\sqrt{s}}} \mathbb{E}\left\{\|\hat{x}_0 - \hat{x}\|^2\right\} + \frac{\sum_{s=1}^{t} \frac{\sigma^2}{s} B^2}{2\sum_{s=1}^{n} \frac{\sigma}{\sqrt{s}}}$$

$$\leq \frac{1}{\sigma\sqrt{n}} \mathbb{E}\left\{\|x_0 - \hat{x}\|^2\right\} + \frac{\sigma B^2}{\sqrt{n}} \underbrace{\sum_{s=1}^{t} \frac{1}{s}}$$

$$\leq \frac{1}{\sqrt{n}} \left(\frac{D^2}{\sigma} + \sigma B^2 (1 + \log(n))\right) \quad (1 + \log(n))$$

Setting $\sigma = \frac{D}{B}$ $\Rightarrow$ Result proved.

Remarks

- Similar proofs for subgradient descent as relying on
$$F(u) \leq F(v) + \eta(u-v) \quad \text{with } \eta \in \partial F(u)$$
- Work with projected gradient descent as using non-expansivity of P_C.
(First steps of the proof).
- convergence of the gradient descent
$$F(x_n) - F(\hat{x}) \leq \frac{L}{2k} \|x_0 - \hat{x}\|^2$$
→ Faster than stochastic.

Convergence of SGD in the strongly convex case

Ass : , F convex , B Lipschitz , $G = F + \frac{\nu}{2}\|\cdot\|_2^2$ admits a unique minimizer $\hat{x}$.
, unbiased gradient
, bounded gradient.

Set $\gamma_k = \frac{1}{\nu k}$, the iterates $(x_k)_{k \geq 0}$ of SGD satisfy.
$$\mathbb{E}\{F(\overline{x_k}) - F(\hat{x})\} \leq \frac{2B^2}{\nu k}(1 + \log k).$$

- Results Bach & Moulines 2011

→ dedicated to stochastic gradient descent with learning rate $\gamma_n = C n^{-\alpha}$

→ strongly smooth case.

- $O(k^{-1})$ for $\alpha = 1$ without averaging
- $O(k^{-1})$ for $\alpha \in [\frac{1}{2}, 1]$ with averaging
- Robust to the choice of C.

5. Stochastic Proximal Gradient [Rosasco - Villa - Vũ 2014]

Ass : $(\gamma_k)_{k \in \mathbb{N}^*}$ positive sequence

$(d_k)_{k \in \mathbb{N}^*}$ a sequence in $[0,1]$

$(G_k)_{k \in \mathbb{N}^*}$ be a $\mathcal{H}$-valued random process s.t $\mathbb{E}\{\|G_k\|^2\} < +\infty$

Fix X_1 a $\mathcal{H}$-valued integrable vector with $\mathbb{E}\{\|X_1\|^2\} < +\infty$

and $\forall k \in \mathbb{N}^*$

$$\begin{cases} z_k = x_k - \gamma_k G_k \\ y_k = \text{prox}_{\gamma_k R}(z_k) \\ \omega_{k+1} = (1-d_k)\, x_k + d_k\, y_k \end{cases}$$

- Si $G_k = \nabla F(\omega_k)$, we get FBS. [Combettes, Wajs 2004]

f, convex, diff, β-Lipschitz continuous gradient. $h \in \Gamma_0(\mathcal{H})$

Ass: $\mathbb{E}\{G_n \mid \mathcal{F}_n\} = \nabla F(x_n)$ ← unbiased estimate of the gradient

weaker condition than $\exists B > 0$ and $\alpha_n > 0$ such that $\mathbb{E}\{\|G_n - \nabla F(x_n)\|^2 \mid \mathcal{F}_n\}$

$\|g_k(x_{k-1})\|_2^2 \leq B^2$ $\leqslant B^2(1 + \alpha_n \|\nabla F(x_n)\|^2$

$\exists \varepsilon > 0$ s.t. $\forall n > 0$ $\quad 0 < \gamma_n < \dfrac{1-\varepsilon}{\beta(1 + 2\sigma^2\alpha_n)}$

Under technical assumptions.

For any solution $\hat{x}$ of the problem $\min f(x) + h(x)$, set

$(\forall k \in \mathbb{N}^*)$ $\chi_k^2 = d_k \gamma_k^2 (1 + 2\alpha_k \|\nabla f(\hat{x})\|^2)$

Assume that $\begin{cases} \sum_{n \in \mathbb{N}^*} d_n \gamma_n = +\infty \\ \sum_{n \in \mathbb{N}^*} \chi_n^2 < +\infty \end{cases}$

$$\mathbb{E}\{\|x_n - \hat{x}\|^2\} = \begin{cases} O(n^{-\theta}) & \text{if } \theta \in]0,1[\\ O(n^{-c}) + O(n^{-1}) & \text{if } \theta = 1 \end{cases}$$

→ Extension de Bach et Moulines 2011.

- Dropout : desactivate neurones output i.e components of $x = (x_j)_{1 \leq j \leq d}$.

It is done randomly.

Each neuron being possibly desactivated during one learning iteration, this forces each unity to correctly learn independently from the others.

It may help to accelerate the learning.