Summary Lecture

V. Estellers

WS 2017

Convexity

Convexity of $E : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$: For all $u, v \in \mathbb{R}^n$ and all $\theta \in [0, 1]$ it holds that

$$E(\theta u + (1-\theta)v) \leq \theta E(u) + (1-\theta)E(v) \tag{c}$$

We call E **strictly convex**, if the inequality (c) is strict for all $\theta \in (0, 1)$, and $v \neq u$.

We call E **m-strongly convex** if $G(u) = E(u) - \frac{m}{2}\|u\|^2$ is convex.

Existence+uniqueness

The **domain** of E is

$$\text{dom}(E) := \{u \in \mathbb{R}^n \mid E(u) < \infty\}.$$

We call E **proper** if $\text{dom}(E) \neq \emptyset$.

The **epigraph** of E is defined as

$$\text{epi}(E) := \{(u, \alpha) \mid E(u) \leq \alpha\}.$$

A function is called **closed** if its epigraph is a closed set.

If E is closed and there exists a nonempty and bounded sublevelset

$$\{u \in \mathbb{R}^n \mid E(u) \leq \alpha\},$$

then E **has a minimizer**.

The subdifferential: Optimality Conditions

The **subdifferential** of a convex function E is

$$\partial E(u) = \{p \in \mathbb{R}^n \mid E(v) - E(u) - \langle p, v - u \rangle \geq 0 \quad \forall v \in \mathbb{R}^n\}$$

If E is differentiable at u then

$$\partial E(u) = \{\nabla E(u)\}.$$

For convex functions, any local minimizer is a global minimizer. The **optimality condition** is

$$\hat{u} \in \arg\min_u E(u) \Leftrightarrow 0 \in \partial E(\hat{u})$$

If E has a minimizer and is strictly convex, the minimizer of E is unique.

The subdifferential: Sum and Chain Rules

The **relative interior** of a convex set M is defined as

$$\mathrm{ri}(M) := \{x \in M \mid \forall y \in M,\ \exists \lambda > 1,\ \text{s.t. } \lambda x + (1-\lambda)y \in M\}.$$

If E is proper and convex and $u \in \mathrm{ri}(\mathrm{dom}(E))$, $\partial E(u)$ **is non-empty**.

Sum rule – Let E_1, E_2 be convex functions such that $\mathrm{ri}(\mathrm{dom}(E_1)) \cap \mathrm{ri}(\mathrm{dom}(E_2)) \neq \emptyset$, then it holds that

$$\partial(E_1 + E_2)(u) = \{p_1 + p_2 \mid p_1 \in \partial E_1(u),\ p_2 \in \partial E_2(u)\}.$$

Chain rule – If $A \in \mathbb{R}^{m \times n}$, $E : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is convex, and $\mathrm{ri}(\mathrm{dom}(E)) \cap \mathrm{range}(A) \neq \emptyset$, then it holds that

$$\partial(E \circ A)(u) = \{A^T p \mid p \in \partial E(Au)\}.$$

Contractions

Question: When does the following **fixed-point iteration** converge?

$$u^{k+1} = G(u^k) \tag{fp}$$

We call $G : \mathbb{R}^n \to \mathbb{R}^n$ a **contraction** if it is Lipschitz-continuous with constant $L < 1$, i.e. if there exists a $L < 1$ such that for all $u,\ v \in \mathbb{R}^n$

$$\|G(u) - G(v)\|_2 \leq L\|u - v\|_2.$$

If G is a contraction, it has a **unique fixed-point** $\hat{u}$ and (fp) **converges linearly** to $\hat{u}$.

Averaged operators

An operator $H : \mathbb{R}^n \to \mathbb{R}^n$ is **non-expasive** if it is Lipschitz-continuous with constant 1, i.e. if for all $u,\ v \in \mathbb{R}^n$

$$\|H(u) - H(v)\|_2 \leq \|u - v\|_2.$$

An operator $G : \mathbb{R}^n \to \mathbb{R}^n$ is called **averaged** if there exists a non-expansive mapping $H : \mathbb{R}^n \to \mathbb{R}^n$ and a constant $\alpha \in (0, 1)$ s.t.

$$G = \alpha I + (1 - \alpha)H.$$

If $G : \mathbb{R}^n \to \mathbb{R}^n$ is averaged and has a fixed-point, then the iteration

$$u^{k+1} = G(u^k)$$

converges to a fixed point of G for any starting point $u^0 \in \mathbb{R}^n$.

Averaged operators

An operator $G : \mathbb{R}^n \to \mathbb{R}^n$ is called **firmly nonexpansive**, if for all $u, v \in \mathbb{R}^n$ it holds that

$$\|G(u) - G(v)\|_2^2 \leq \langle G(u) - G(v), u - v \rangle.$$

An operator $G : \mathbb{R}^n \to \mathbb{R}^n$ is **firmly nonexpansive** if and only if G is **averaged with** $\alpha = \frac{1}{2}$.

Compositions of averaged operators are averaged.

Gradient descent

Gradient descent iteration: $u^{k+1} = u^k - \tau \nabla E(u^k)$

E is L-**smooth** if E is differentiable and ∇E is L-Lipschitz continuous.

Baillon-Haddad Th.: A continuously differentiable convex function $E : \mathbb{R}^n \to \mathbb{R}$ is L-smooth if and only if $\frac{1}{L}\nabla E$ is firmly nonexpansive.

Theorem: For E convex and L-smooth, gradient descent with a fixed step-size $\tau \in (0, 2/L)$ converges to a solution of $\min_{u \in \mathbb{R}^n} \; E(u)$.

As E is L-smooth, $\frac{1}{L}\nabla E = \frac{1}{2}(I + T)$ for some non-expansive operator T.

$$G(u) = u - \tau L \frac{1}{L} \nabla E(u) = \left(1 - \frac{L\tau}{2}\right) I + \frac{L\tau}{2}(-T)$$

is averaged for $\tau \in (0, 2/L)$. Then $u^{k+1} = G(u^k) = u^k - \tau \nabla E(u^k)$ converges to a fixed-point of G (a minimizer of E as $\nabla E(u^*) = 0$).

Gradient projection

Consider the problem $\min_{u \in C} \; E(u)$ for $E, \; C$ convex and E L-smooth.

The **gradient projection iteration** is $u^{k+1} = \text{proj}_C(\underbrace{u^k - \tau \nabla E(u^k)}_{G_\tau(u^k)})$.

We can sow that the projection onto a non-empty closed convex set is firmly nonexpansive $\Rightarrow$ proj_C is averaged.
If E is L-**smooth** and $\tau \in (0, 2/L)$, then G_τ is averaged and $\text{proj}_C(G_\tau)$ is averaged because the composition of averaged operators is averaged.
Then the gradient projection alg. converges to a minimizer of E over C as a fixed-point iteration $u^{k+1} = \text{proj}_C(G_\tau(u^k))$ of an averaged operator.

Proximal Operator

The mapping $\text{prox}_E : \mathbb{R}^n \to \mathbb{R}^n$ defined as

$$\text{prox}_E(v) := \operatorname*{argmin}_{u \in \mathbb{R}^n} \; E(u) + \frac{1}{2} \|u - v\|^2$$

for a closed, proper, convex function $E : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is called the **proximal operator** or **proximal mapping** of E.

The proximal operator prox_E for a closed, proper, convex function E is **firmly nonexpansive** and therefore **averaged with** $\alpha = 1/2$.

Proximal gradient

Consider the problem $\min_{u\in\mathbb{R}^n}\ F(u)+G(u)$ for F convex and G convex and L smooth. Then the iteration

$$u^{k+1} = \mathsf{prox}_{\tau F}(u^k - \tau\nabla G(u^k))$$

is called the **proximal gradient method**.

Let $E(u) = F(u) + G(u)$ **have a minimizer**, and $\tau \in (0, 2/L)$, then the proximal gradient method **converges** to a minimizer of E.

The **convergence rates** of gradient descent, gradient projection, and proximal gradient are **suboptimal**. They are accelerated by extrapolation.

Convex conjugation

The **convex conjugate** of a proper function $E : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is

$$E^*(p) = \sup_u \langle u, p \rangle - E(u).$$

It is always convex and closed.

The **Fenchel-Young inequality** states that

$$E(u) + E^*(p) \geq \langle u, p \rangle,$$

and that equality holds if and only if $p \in \partial E(u)$.

For a proper, closed convex function E, its biconjugate $E^{**} = E$.

For a proper, closed convex function E, $p \in \partial E(u) \Leftrightarrow u \in \partial E^*(p)$.

Fenchel Duality

Let $E(u) = G(u) + F(Ku)$ have a minimizer, and let G and F be closed and convex. If there is $u \in \text{ri}(\text{dom}(G))$ s.t. $Ku \in \text{ri}(\text{dom}(F))$, then

$$
\boxed{
\begin{array}{llll}
 & \min_u & G(u) + F(Ku) & \textbf{Primal} \\[2ex]
= & \min_u \max_q & G(u) + \langle q, Ku \rangle - F^*(q) & \\
 & & & \textbf{Saddle point} \\
= & \max_q \min_u & G(u) + \langle q, Ku \rangle - F^*(q) & \\[2ex]
= & \max_q & -G^*(-K^* q) - F^*(q) & \textbf{Dual}
\end{array}
}
$$

We are therefore looking for a **saddle point** (u, q) such that

$$-K^T q \in \partial G(u), \quad Ku \in \partial F^*(q).$$

PDHG

The primal-dual view motivates the definition of an iterative method to find

$$-K^T q \in \partial G(u), \quad Ku \in \partial F^*(q).$$

The **primal-dual hybrid gradient (PDHG)** method computes

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \underbrace{\begin{pmatrix} \partial G & K^T \\ -K & \partial F^* \end{pmatrix}}_{=:T} \begin{pmatrix} u^{k+1} \\ p^{k+1} \end{pmatrix} + \underbrace{\begin{pmatrix} \frac{1}{\tau} I & -K^T \\ -K & \frac{1}{\sigma} I \end{pmatrix}}_{=:M} \begin{pmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{pmatrix}.$$

or in the **algorithmic-friendly form** of (PDHG)

$$\begin{aligned} p^{k+1} &= \mathsf{prox}_{\sigma F^*}(p^k + \sigma K(2u^k - u^{k-1})), \\ u^{k+1} &= \mathsf{prox}_{\tau G}(u^k - \tau K^* p^{k+1}), \end{aligned} \tag{PDHG}$$

Convergence analysis

A set-valued operator T is called **monotone** (a generalization of firmly non-expansive) if $\langle p-q, u-v \rangle \geq 0 \qquad \forall u, v, p \in T(u), q \in T(v)$.

The **resolvent** $(I+T)^{-1}$ of a maximally monotone operator is firmly non-expansive, i.e. **averaged with** $\alpha = 1/2$.

Let T be maximally monotone and let there exist a z such that $0 \in T(z)$. Then the **proximal point algorithm**

$$0 \in T(z^{k+1}) + z^{k+1} - z^k$$

converges to a $\tilde{z}$ with $0 \in T(\tilde{z})$.

Convergence of PDHG

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \underbrace{\begin{pmatrix} \partial G & K^T \\ -K & \partial F^* \end{pmatrix}}_{=:T} \begin{pmatrix} u^{k+1} \\ p^{k+1} \end{pmatrix} + \underbrace{\begin{pmatrix} \frac{1}{\tau} I & -K^T \\ -K & \frac{1}{\sigma} I \end{pmatrix}}_{=:M} \begin{pmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{pmatrix}.$$

T is **maximally monotone**, M is **positive definite** for $\tau\sigma < \frac{1}{\|K\|_{S^\infty}^2}$.

Let $M = M^{1/2} M^{1/2}$, then $M^{-1/2} T M^{-1/2}$ is maximally montone and the **PDHG algorithm is a proximal point algorithm** in $z = M^{1/2}(u; p)$.

If saddle-point problem has a solution and $\tau\sigma < \|K\|_{S^\infty}^{-2}$, PDHG converges.

PDHG

The **variants of PDHG** for functions F^* or G **strongly convex converge faster**.

Considering

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \partial G & K^T \\ -K & \partial F^* \end{pmatrix} \begin{pmatrix} u^{k+1} \\ p^{k+1} \end{pmatrix} + \begin{pmatrix} \frac{1}{\tau} I & -K^T \\ -K & \frac{1}{\sigma} I \end{pmatrix} \begin{pmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{pmatrix}.$$

we measure convergence, and define stopping criteria, in terms of the residuals

$$r_p^{k+1} = \frac{1}{\sigma}(p^{k+1} - p^k) - K(u^{k+1} - u^k)$$
$$r_d^{k+1} = \frac{1}{\tau}(u^{k+1} - u^k) - K^T(p^{k+1} - p^k)$$

Summary: Learning Problem

Given a set of examples $(x_1, y_1), \ldots, (x_n, y_n)$

each example $\xi = (x, y)$ is a pair of an input x and a scalar output y.

loss $\ell(\hat{y}, y)$ measures the cost of predicting $\hat{y}$ when the answer is y

family of functions $h(\cdot; w)$ parametrized by a weight vector w.

We seek $h \in$ that minimizes the loss $f(\xi; w) = \ell(h(x; w), y)$.
Although we would like to average over the unknown distribution $P(x, y)$

$$f(w) = R(w) = \mathbb{E}[\ell(h(x; w), y)] = \int \ell(h(x; w), y) \mathrm{d}P(x, y)$$

we must settle for computing the average over the samples

$$f(w) = R_n(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i; w), y_i).$$

Statistical learning theory (Vapnik and Chervonenkis, 1971) justifies minimizing R_n instead of R when is sufficiently restrictive.

Stochastic Gradient Method

The objective function $F\colon \mathbb{R}^d \mapsto \mathbb{R}$ can be the expected or empirical risk:

$$F(w) = \mathbb{E}[f(w,\xi)] \qquad \text{or} \qquad F(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w).$$

The analysis applies to both objectives, depending on how the stochastic gradient estimates are chosen.

Stochastic Gradient Method [0] Choose an initial iterate w_1 k=1,2,... Generate a realization of the random variable ξ_k Compute a stochastic vector $g(w_k,\xi_k)$ Choose a stepsize $\alpha_k > 0$ Set the new iterate as

$$w_{k+1} = w_k - \alpha_k g(w_k,\xi_k)$$

Fundamental Lemmas

Lemma

If F is an L-smooth function, the iterates of SG satisfy:

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \underbrace{\nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k,\xi_k)]}_{\substack{\textit{expected directional derivative} \\ \textit{of } F \textit{ along direction } g(w_k,\xi_k)}} + \frac{\alpha_k^2 L}{2} \underbrace{\mathbb{E}_{\xi_k}[\|g(w_k,\xi_k)\|^2]}_{\textit{second moment } g(w_k,\xi_k)}$$

Lemma

If F is L-smooth and there are $M \leq 0$ and $M_G \geq \mu^2 \geq 0$ such that

$$\mathbb{E}_{\xi_k}[\|g(w_k,\xi_k)\|^2] \leq M + M_G \|\nabla F(w_k)\|^2,$$

then the SG iterates satisfy

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\underbrace{(\mu - \frac{1}{2}\alpha_k L M_G)\alpha_k \|\nabla F(w_k)\|^2 + \frac{1}{2}\alpha_k^2 L M}_{\textit{deterministic}}.$$

Convergence of SG

Theorem

If F is L-smooth and c-strongly convex and satisfies Assumption 2, then the SG method run with a positive stepsize $\alpha \leq \frac{\mu}{LM_G}$ satisfies

$$\mathbb{E}[F(w_k) - F^*] \leq \frac{\alpha LM}{2c\mu} + (1 - \alpha c\mu)^{k-1}\left(F(w_1) - F^* - \frac{\alpha LM}{2c\mu}\right),$$

Theorem

If F is L-smooth and c-strongly convex and satisfies Assumption 2, then SG method with stepsizes $\alpha_k = \frac{\beta}{\gamma+k}$ for some $\beta > \frac{1}{c\mu}$, $\gamma > 0$ such that $\alpha_1 \leq \frac{\mu}{LM_G}$ satisfies

$$\mathbb{E}[F(w_k) - F^*] \leq \frac{\eta}{\gamma + k} \qquad \eta = \max\left\{\frac{\beta^2 LM}{2(\beta c\mu - 1)}, (\gamma + 1)(F(w_1) - F^*)\right\}.$$

Noise-Reduction Methods

Noise-Reduction Methods: instead of decreasing the learning rate to converge to the optimum, reduce variance of the stochastic gradients. They achieve a linear convergence rate at a higher per-iteration cost.

Other methods come with few guarantees but work well in practice:

- Gradient Methods with Momentum
- Accelerated Gradient Method
- Adaptive Methods: adagrad, adadelta, adam

Primal-Dual Algorithms

Make sure the updates decouple, are easy, and $M \succeq 0$

PDHG, overrelaxation on primal

$$0 \in \begin{bmatrix} \partial G & K^T \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} + \begin{bmatrix} \frac{1}{\tau} I & -K^T \\ -K & \frac{1}{\sigma} I \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix}.$$

PDHG, overrelaxation on dual

$$0 \in \begin{bmatrix} \partial G & K^T \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} + \begin{bmatrix} \frac{1}{\tau} I & K^T \\ K & \frac{1}{\sigma} I \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix}.$$

Primal ADMM

$$0 \in \begin{bmatrix} \partial G & K^T \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} + \begin{bmatrix} \lambda K^T K & K^T \\ K & \frac{1}{\lambda} I \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix}.$$

Corresponding dual ADMM

$$0 \in \begin{bmatrix} \partial G & K^T \\ -K & \partial F^* \end{bmatrix} \begin{bmatrix} u^{k+1} \\ p^{k+1} \end{bmatrix} + \begin{bmatrix} \frac{1}{\lambda} I & -K^T \\ -K & \lambda K K^T \end{bmatrix} \begin{bmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{bmatrix}.$$