Mirror descent, Hedge

October 17, 2018

Gradient descent

- Goal : minimize $f : \mathbb{R}^n \to \mathbb{R}$
- Gradient descent :
  Suppose f is convex and differentiable. from a given point x_0, generate sequence
  $$x_{k+1} := x_k - \lambda_k f'(x_k)$$
  where $\lambda_k > 0$ and $f'(x_k)$ is the vector for which
  $$f'(x_k)^\top h = \lim_{\lambda \searrow 0} \frac{f(x_k + \lambda h) - f(x_k)}{t} \qquad \forall h \in \mathbb{R}^n$$

Subgradient descent

- Suppose f is convex and not differentiable, but closed (its epigraph is closed). then we can define its subgradient g_k as a substitute for the gradient $f'(x_k)$, that is an element of the set:

$$\partial f(x_k) := \{g \in \mathbb{R}^n : f(x) \geq f(x_k) + (x - x_k)^\top g \text{ for all } x \in \mathbb{R}^n\}.$$

- Definition depends on a scalar product that we have chosen arbitrarily.

Generalized subgradient

- Generalization of subgradient descent (Nemirovski and Yudin, 1979)
- Let E be a Euclidean space and E^* is its dual, i.e., the set of all linear applications from E to $\mathbb{R}$. For $h \in E^*$ and $x \in E$, Denote $< h, x >= h(x)$.
- Denote $||\cdot||$ a norm on E. Then its corrensponding norm on E^* is defined as

$$||h||_* := \max_{||x||=1, x \in E} < h, x >, \qquad h \in E^*$$

Then the definition of subgradient can easily extend to more general setting:

$$\partial f(x_k) := \{g \in E^* : f(x) \geq f(x_k) + < g, x - x_k > \text{ for all } x \in E\}.$$

Mirror descent

- Goal : solve convex optimization problems that are formulated as:

$$f^* := \min_{x \in Q} f(x)$$

where $Q \subseteq E$ is a closed convex set, The function $f : Q \to \mathbb{R}$ is closed, convex, and equipped with an *oracle*, which means for all $x \in Q$ we can compute the value $f(x)$ and its subgradient $g \in \partial f(x)$.

- **Mirror descent algorithm**
Let V_Q be a map from E^* to Q. Set $s_0 := 0$ and select a set of step-sizes $\{\lambda_k\}_{k \geq 0}$ and a starting point $x_0 \in Q$.

For $k = 0, 1, \cdots,$

1. Determine $g_k \in \partial f(x_k)$.
2. set $s_{k+1} := s_k - \lambda_k g_k$.
3. compute $x_{k+1} := V_Q(s_{k+1})$.

Constuction of V_Q

- Mirror descent algorithm requires a *prox-function* $d : Q \to \mathbb{R}$, that is, strongly convex continuously differentiable function: there exist $\sigma > 0$ such that for every $x, y \in Q$,

$$d(y) \geq d(x) + < d'(x), y - x > + \frac{\sigma}{2} ||y - x||^2.$$

  And, assume that d has a (unique) minimizer x_0 on Q.

- Define V_Q as

$$V_Q(s) := \operatorname*{argmax}_{x \in Q} \{ < s, x - x_0 > - d(x) \}$$

  It is well-defined since d is strongly convex.

Convergence for Mirror descent

- **Theorem**
  Assume that there exist a constant D s.t. $D \geq d(x^*)$, where $x^* \in Q$ and $f(x^*) = f^*$). With $f_k := \min\{f(x_i) : 0 \leq i \leq k\}$, we have:

$$f_k - f^* \leq \frac{1}{\sum_{i=0}^{k} \lambda_i} \left(D + \frac{1}{2\sigma} \sum_{i=0}^{k} \lambda_i^2 ||g_i||_*^2 \right).$$

- If there is a constant Γ for which $||g_i||_* \leq \Gamma$ for all i, Then the above algorithm is guaranteed to converge as long as $\sum_{i=0}^{k} \lambda_i$ diverges and $\sum_{i=0}^{k} \lambda_i^2$ converges as k goes to infinity.
- The later condition implies that $\lim_{k\to\infty} \lambda_k = 0$.
- (?) new subgradients should be treated with more consideration than old ones as they are likely to contain more relevant information.

Nesterov's Primal-Dual Subgradient Algorithm

- Given a paramtere $\beta > 0$, we set:

$$V_{Q,\beta}(s) := \underset{x \in Q}{\text{argmax}}\{< s, x - x_0 > -\beta d(x)\}$$

- **Nesterov's algorithm**
  Set $s_0 := 0$, select a set of step-sizes $\{\lambda_k\}_{k \geq 0}$ and a non-decreasing sequence $\{\beta_k\}_{k \geq 0}$ of projection paramters. Set $x_0 := \text{argmin}\{d(x) : x \in Q\}$.

  For $k = 0, 1, \cdots,$

  1. Determine $g_k \in \partial f(x_k)$.
  2. set $s_{k+1} := s_k - \lambda_k g_k$.
  3. compute $x_{k+1} := V_{Q,\beta_{k+1}}(s_{k+1})$.

Convergence for Nesterov's algorithm

- Define a *regret* R_k as:

$$R_k := \max\left\{\sum_{i=0}^{k} \lambda_i < g_i, x_i - x >: x \in Q, d(x) \leq D\right\}.$$

- **Theorem**
Assume that there exist a constant D s.t. $D \geq d(x^*)$, where $x^* \in Q$ and $f(x^*) = f^*)$. With $f_k := \min\{f(x_i) : 0 \leq i \leq k\}$, we have:

$$f_k - f^* \leq \frac{R_k}{\sum_{i=0}^{k} \lambda_i} \leq \frac{1}{\sum_{i=0}^{k} \lambda_i}\left(\beta_{k+1} D + \frac{1}{2\sigma}\sum_{i=0}^{k}\frac{\lambda_i^2}{\beta_i}||g_i||_*^2\right).$$

- If we choose $\lambda_j = 1$ for all j, $\beta_{j+1} := \nu\hat{\beta}_{j+1}$, $\hat{\beta}_0 = 1$ and $\hat{be}_{j+1} = \sum_{i=0}^{j} \frac{1}{\hat{\beta}_i}$,
and $\nu := \frac{\Gamma}{\sqrt{2\sigma D}}$, RHS $= O(k^{-0.5})$

Stochastic descent

- Goal : Given a Borel probability space $(\Omega, \mathcal{B}, P)$ and an objective function $\phi : Q \times \Omega \to \mathbb{R}$ (loss function) that is P- integrable for each fixed x and where $Q \subseteq E$ is the feasible set, we aim at solving:

$$f^* := \min_{x \in Q} E_P[\phi(x, \omega)] = \min_{x \in Q} f(x).$$

- However, we don't know about P, so we can't compute and value about f. Instead, we observe a series of samples $\{\omega_{k,\alpha}\}_{1 \leq \alpha \leq L_k} \subseteq \Omega$.
- Instead of using $g_k \in \partial f(x_k)$, use stochasitc subgradient of f at x_k, $\tilde{g}_k := \sum_{\alpha=1}^{L_k} \nabla_x \phi(x_k, \omega_{k,\alpha})/L_k$, where $\nabla_x \phi(x_k, \omega_{k,\alpha}) \in \partial_x \phi(x_k, \omega_{k,\alpha})$.

Stochastic Mirror descent

- **Theorem**
  Suppose we use $\tilde{g}_k$ instead of g_k in Nesterov's algorithm. Let $M_k := \sum_{i=0}^{k}$ and
  $$\tilde{f}_k := \min_{0 \leq i \leq k} E_{P^{M_k}}\left[\phi(x_i, \omega)\right]$$
  , we have:
  $$\tilde{f}_k - f^* \leq \frac{1}{\sum_{i=0}^{k} \lambda_i} \left(\beta_{k+1} D + \frac{1}{2\sigma} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_i} ||g_i||_*^2 \right).$$

Stochastic Mirror descent

- **Theorem**
  Assume that the above conditions hold and let

$$V := max\{\phi(x,\omega) - \phi(x,\omega') : \omega, \omega' \in \Omega, x \in Q\} < \infty.$$

For every $\epsilon > 0$, the inequality

$$\min_{0 \le i \le k} f(x_i) - f^* \le \frac{1}{\sum_{i=0}^{k} \lambda_i} \left(\beta_{k+1} D + \frac{1}{2\sigma} \sum_{i=0}^{k} \frac{\lambda_i^2}{\beta_i} \Gamma^2 \right) + 2\epsilon.$$

hols with a probability of at least

$$1 - 2\exp\left(- \frac{2\epsilon^2 (\sum_{j=0}^{k} \lambda_j)^2}{M_k V^2} \min_{0 \le i \le k} \frac{L_i^2}{\lambda_i^2} \right).$$

On-line allocation model

- The allocation agent A has N options or stratagies to choose from. At each time step $k = 1, 2, \cdots, T$
- The allocator A decides on a distribution p^k over the strategies; $p_i^k \geq 0$ is the amount allocated to strategy i, and $\sum_{i=1}^{N} p_i^k = 1$.
- Each strategy i then suffers some loss l_i^t which is determined by the 'enviroment'.
- Let $\omega_k \in \Omega$ be a k-th sample. The loss suffered by A is then $\sum_{i=1}^{n} p_i^k l_i(\omega_k) = p^k \cdot l^k$, i.e. the average loss of the strategies w.r.t. A's chosen allocation rule.

Hedge Algorithm

- Motive : Littlestone and Warmuth's weighted majority algorithm (1994)
- Parameters : $\beta \in [0,1]$
- Choose utility function $U_\beta : [0,1] \to [0,1]$ satisfying

$$\beta^r \leq U_\beta(r) \leq 1 - (1-\beta)r$$

  for all $r \in [0,1]$. General choice is $U_\beta(r) = \beta^r$.
- Let $w^k = (w_1^k, \cdots, w_N^k)^\top$ be a weight vector (unnormalized ι^k). Initialize w^1.
  For $k = 1, 2, \cdots, T$,
  1. Choose allocation

$$p^k = \frac{w^k}{\sum_{i=1}^N w_i^k}$$

  2. Draw $\omega_k \in \Omega$
  3. Receive loss vector $l(\omega_k) \in [0,1]^N$.
  4. Update weight as

$$w_i^{k+1} = w_i^k \cdot U_\beta(l_i^k)$$

  At each iteration k, our strategy faces a loss of $\mathcal{L}_k := \sum_{i=1}^N p_i^k l_i(\omega_k)$

Hedge Algorithm

- **Theorem**
  For some $\beta \in [0,1]$, we have

$$\sum_{k=0}^{T} \mathcal{L}_k - \min_{1 \le i \le N} \sum_{k=1}^{T} l_i(\omega_k) \le \sqrt{2T \log N} + \log N.$$

  Therefore, if T increases as infinity, average regret converges to 0.

Hedge Algorithm as Stochastic Mirror descent algorithm

- For $U_\beta : [0,1] \to [0,1]$, Let $\phi(x,\omega) = -\sum_{i=1}^{N} x_i \log U_\beta(l_i(\omega))$. and consider the optimization problem

$$\min_{x \in S_n} E_P[\phi(x,\omega)]$$

where the set S_n is the standard simplex of $\mathbb{R}^N$:

$$S_n := \{x \in \mathbb{R}_+^N : \sum_{i=1}^{N} x_i = 1\}$$

Then, $\nabla_x \phi(x,\omega) = -log U_\beta(l(\omega))$.

Hedge Algorithm as Stochastic Mirror descent algorithm

- If we use prox-function d as entropy function :

$$d : S_n \to \mathbb{R}, \qquad x \to d(x) := \sum_{i=1}^{N} x_i \log x_i + \log N$$

, Then corresponding mirror operator takes the following form :

$$V_Q(s) = \text{argmax}\{< s, x - x_0 > -d(x) : x \in S_n\} = \left[\frac{\exp(s_i)}{\sum_{j=1}^{n} \exp(s_j)} \right]_{1 \le i \le n}$$

Then, the solution of stochastic mirror descent algorithm is equal to the solution of Hedge algorithm.