18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet
Scribe: Sylvain Carpentier

Lecture 14
Oct. 26, 2015

In this lecture we will wrap up the study of optimization techniques with stochastic optimization. The tools that we are going to develop will turn out to be very efficient in minimizing the φ-risk when we can bound the noise on the gradient.

3. STOCHASTIC OPTIMIZATION

3.1 Stochastic convex optimization

We are considering random functions $x \mapsto \ell(x, Z)$ where x is the optimization parameter and Z a random variable. Let P_Z be the distribution of Z and let us assume that $x \mapsto \ell(x, Z)$ is convex P_Z a.s. In particular, $\mathbb{E}[\ell(x, Z)]$ will also be convex. The goal of stochastic convex optimization is to approach $\min_{x \in \mathcal{C}} \mathbb{E}[\ell(x, Z)]$ when $\mathcal{C}$ is convex. For our purposes, $\mathcal{C}$ will be a deterministic convex set. However, stochastic convex optimization can be defined more broadly. The constraint can be itself stochastic :

$$\mathcal{C} = \{x, \mathbb{E}[g(x, Z)] \leq 0\}, \quad g \text{ convex } P_Z \text{ a.s.}$$

$$\mathcal{C} = \{x, \mathbb{P}[g(x, Z) \leq 0] \geq 1 - \varepsilon\}, \quad \text{“chance constraint”}$$

The second constraint is not convex a priori but remedies are possible (see [NS06, Nem12]). In the following, we will stick to the case where X is deterministic. A few optimization problems we tackled can be interpreted in this new framework.

3.1.1 Examples

Boosting. Recall that the goal in Boosting is to minimize the φ-risk:

$$\min_{\alpha \in \Lambda} \mathbb{E}[\varphi(-Y f_\alpha(X))]\,,$$

where Λ is the simplex of $\mathbb{R}^d$. Define $Z = (X, Y)$ and the random function $\ell(\alpha, Z) = \varphi(-Y f_\alpha(X))$, convex P_Z a.s.

Linear regression. Here the goal is the minimize the ℓ_2 risk:

$$\min_{\alpha \in \mathbb{R}^d} \mathbb{E}[(Y - f_\alpha(X))^2]\,.$$

Define $Z = (X, Y)$ and the random function $\ell(\alpha, Z) = (Y - f_\alpha(X))^2$, convex P_Z a.s.

Maximum likelihood. We consider samples $Z_1, \ldots, Z_n$ iid with density p_θ, $\theta \in \Theta$. For instance, $Z \sim \mathcal{N}(\theta, 1)$. The likelihood functions associated to this set of samples is $\theta \mapsto \prod_{i=1}^n p_\theta(Z_i)$. Let $p^*(Z)$ denote the true density of Z (it does not have to be of the form p_θ for some $\theta \in \Theta$. Then

$$\frac{1}{n}\mathbb{E}[\log \prod_{i=1}^n p_\theta(Z_i)] = -\int \log(\frac{p^*(z)}{p_\theta(z)}) p^*(z) dz + C = -\mathsf{KL}(p^*, p_\theta) + C$$

where C is a constant in θ. Hence maximizing the expected log-likelihood is equivalent to minimizing the expected Kullback-Leibler divergence:

$$\max_\theta \mathbb{E}[\log \prod_{i=1}^n p_\theta(Z_i)] \iff \mathsf{KL}(p^*, p_\theta)$$

External randomization. Assume that we want to minimize a function of the form

$$f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)\,,$$

where the functions $f_1, \ldots, f_n$ are convex. As we have seen, this arises a lot in empirical risk minimization. In this case, we treat this problem as deterministic problem but inject artificial randomness as follows. Let I be a random variable uniformly distributed on $[n] =: \{1, \ldots, n\}$. We have the representation $f(x) = \mathbb{E}[f_I(x)]$, which falls into the context of stochastic convex optimization with $Z = I$ and $\ell(x, I) = f_I(x)$.

Important Remark: There is a key difference between the case where we assume that we are given independent random variables and the case where we generate artificial randomness. Let us illustrate this difference for Boosting. We are given $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d from some unknown distribution. In the first example, our aim is to minimize $\mathbb{E}[\varphi(-Yf_\alpha(X))]$ based on these n observations and we will that the stochastic gradient allows to do that by take one pair (X_i, Y_i) in each iteration. In particular, we can use each pair at most once. We say that we do *one pass* on the data.

We could also leverage our statistical analysis of the empirical risk minimizer from previous lectures and try to minimize the empirical φ-risk

$$\hat{R}_{n,\varphi}(f_\alpha) = \frac{1}{n}\sum_{i=1}^n \varphi(-Y_i f_\alpha(X_i))$$

by generating k independent random variables $I_1, \ldots, I_k$ uniform over $[n]$ and run the stochastic gradient descent to us one random variable I_j in each iteration. The difference here is that k can be arbitrary large, regardless of the number n of observations (we make *multiple passes* on the data). However, minimizing $\mathbb{E}_I[\varphi(-Y_I f_\alpha(X_I))|X_1, Y_1, \ldots, X_n, Y_n]$ will perform no better than the empirical risk minimizer whose statistical performance is limited by the number n of observations.

3.2 Stochastic gradient descent

If the distribution of Z was known, then the function $x \mapsto \mathbb{E}[\ell(x, Z)]$ would be known and we could apply gradient descent, projected gradient descent or any other optimization tool seen before in the deterministic setup. However this is not the case in reality where the true distribution P_Z is unknown and we are only given the samples $Z_1, \ldots, Z_n$ and the random function $\ell(x, Z)$. In what follows, we denote by $\partial\ell(x, Z)$ the set of subgradients of the function $y \mapsto \ell(y, Z)$ at point x.

Algorithm 1 Stochastic Gradient Descent algorithm

```
Input: $x_1 \in \mathcal{C}$, positive sequence $\{\eta_s\}_{s\geq 1}$, independent random variables $Z_1, \dots, Z_k$
with distribution $P_Z$.
for $s = 1$ to $k-1$ do
  $y_{s+1} = x_s - \eta_s \tilde{g}_s\,, \quad \tilde{g}_s \in \partial \ell(x_s, Z_s)$
  $x_{s+1} = \pi_{\mathcal{C}}(y_{s+1})$
end for
return $\bar{x}_k = \frac{1}{k}\sum_{s=1}^{k} x_s$
```

Note the difference here with the deterministic gradient descent which returns either $\bar{x}_k$ or $x_k^\circ = \underset{x_1,\dots,x_n}{\operatorname{argmin}} f(x)$. In the stochastic framework, the function $f(x) = \mathbb{E}[\ell(x,\xi)]$ is typically unknown and $\overset{\circ}{x}_k$ cannot be computed.

Theorem: Let $\mathcal{C}$ be a closed convex subset of $\mathbb{R}^d$ such that $\operatorname{diam}(\mathcal{C}) \leq R$. Assume that he convex function $f(x) = \mathbb{E}[\ell(x, Z)]$ attains its minimum on $\mathcal{C}$ at $x^* \in \mathbb{R}^d$. Assume that $\ell(x, Z)$ is convex P_Z a.s. and that $\mathbb{E}\|\tilde{g}\|^2 \leq L^2$ for all $\tilde{g} \in \partial \ell(x, Z)$ for all x. Then if $\eta_s \equiv \eta = \frac{R}{L\sqrt{k}}$,

$$\mathbb{E}[f(\bar{x}_k)] - f(x^*) \leq \frac{LR}{\sqrt{k}}$$

Proof.

$$\begin{aligned}
f(x_s) - f(x^*) &\leq {g_s}^\top (x_s - x^*) \\
&= \mathbb{E}[\tilde{g}_s^\top (x_s - x^*)|x_s] \\
&= \frac{1}{\eta}\mathbb{E}[(y_{s+1} - x_s)^\top (x_s - x^*)|x_s] \\
&= \frac{1}{2\eta}\mathbb{E}[\|x_s - y_{s+1}\|^2 + \|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2|x_s] \\
&\leq \frac{1}{2\eta}(\eta^2 \mathbb{E}[\|\tilde{g}_s\|^2|x_s] + \mathbb{E}[\|x_s - x^*\|^2|x_s] - \mathbb{E}[\|x_{s+1} - x^*\|^2|x_s]
\end{aligned}$$

Taking expectations and summing over s we get

$$\frac{1}{k}\sum_{s=1}^{k} f(x_s) - f(x^*) \leq \frac{\eta L^2}{2} + \frac{R^2}{2\eta k}\,.$$

Using Jensen's inequality and chosing $\eta = \frac{R}{L\sqrt{k}}$, we get

$$\mathbb{E}[f(\bar{x_k})] - f(x^*) \leq \frac{LR}{\sqrt{k}}$$

□

3.3 Stochastic Mirror Descent

We can also extend the Mirror Descent to a stochastic version as follows.

Algorithm 2 Mirror Descent algorithm

```
**Input:** $x_1 \in \operatorname{argmin}_{\mathcal{C}\cap\mathcal{D}} \Phi(x)$, $\zeta : \mathbb{R}^d \to \mathbb{R}^d$ such that $\zeta(x) = \nabla\Phi(x)$, independent random variables $Z_1, \dots, Z_k$ with distribution $P_Z$.
**for** $s = 1, \cdots, k$ **do**
    $\zeta(y_{s+1}) = \zeta(x_s) - \eta\tilde{g}_s$ for $\tilde{g}_s \in \partial\ell(x_s, Z_s)$
    $x_{s+1} = \Pi^{\Phi}_{\mathcal{C}}(y_{s+1})$
**end for**
**return** $\overline{x} = \frac{1}{k}\sum_{s=1}^{k} x_s$
```

Theorem: Assume that Φ is α-strongly convex on $\mathcal{C} \cap\ \mathcal{D}$ w.r.t. $\|\cdot\|$ and

$$R^2 = \sup_{x\in\mathcal{C}\cap\mathcal{D}} \Phi(x) - \min_{x\in\mathcal{C}\cap\mathcal{D}} \Phi(x)$$

take $x_1 = \operatorname{argmin}_{x\in\mathcal{C}\cap\mathcal{D}} \Phi(x)$ (assume that it exists). Then, Stochastic Mirror Descent with $\eta = \frac{R}{L}\sqrt{\frac{2\alpha}{R}}$ outputs $\bar{x}_k$, such that

$$\mathbb{E}[f(\bar{x}_k)] - f(x^*) \leq RL\sqrt{\frac{2}{\alpha k}}\,.$$

Proof. We essentially reproduce the proof for the Mirror Descent algorithm.

Take $x^\sharp \in \mathcal{C} \cap \mathcal{D}$. We have

$$\begin{aligned}
f(x_s) - f(x^\sharp) &\leq g_s^\top(x_s - x^\sharp) \\
&\mathbb{E}[\tilde{g}_s^\top(x_s - x^*)|x_s] \\
&= \frac{1}{\eta}\mathbb{E}[(\zeta(x_s) - \zeta(y_{s+1}))^\top(x_s - x^\sharp)|x_s] \\
&= \frac{1}{\eta}\mathbb{E}[(\nabla\Phi(x_s) - \nabla\Phi(y_{s+1}))^\top(x_s - x^\sharp)|x_s] \\
&= \frac{1}{\eta}\mathbb{E}\left[D_\Phi(x_s, y_{s+1}) + D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, y_{s+1})\middle|x_s\right] \\
&\leq \frac{1}{\eta}\mathbb{E}\left[D_\Phi(x_s, y_{s+1}) + D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, x_{s+1})\middle|x_s\right] \\
&\leq \frac{\eta}{2\alpha^2}\mathbb{E}[\|\tilde{g}_s\|_*^2|x_s] + \frac{1}{\eta}\mathbb{E}\left[D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, x_{s+1})\middle|x_s\right]
\end{aligned}$$

where the last inequality comes from

$$
\begin{aligned}
D_\Phi(x_s, y_{s+1}) &= \Phi(x_s) - \Phi(y_{s+1}) - \nabla\Phi(y_{s+1})^\top (x_s - y_{s+1}) \\
&\leq \left[\nabla\Phi(x_s) - \nabla\Phi(y_{s+1})\right]^\top (x_s - y_{s+1}) - \frac{\alpha}{2}\|y_{s+1} - x_s\|^2 \\
&\leq \eta\|\tilde{g}_s\|_*\|x_s - y_{s+1}\| - \frac{\alpha}{2}\|y_{s+1} - x_s\|^2 \\
&\leq \frac{\eta^2\|\tilde{g}_s\|_*^2}{2\alpha}\,.
\end{aligned}
$$

Summing and taking expectations, we get

$$
\frac{1}{k}\sum_{s=1}^{k}[f(x_s) - f(x^\sharp)] \leq \frac{\eta L^2}{2\alpha} + \frac{D_\Phi(x^\sharp, x_1)}{k\eta}\,. \tag{3.1}
$$

We conclude as in the previous lecture.

□

3.4 Stochastic coordinate descent

Let f be a convex L-Lipschitz and differentiable function on $\mathbb{R}^d$. Let us denote by $\nabla_i f$ the partial derivative of f in the direction e_i. One drawback of the Gradient Descent Algorithm is that at each step one has to update every coordinate $\nabla_i f$ of the gradient. The idea of the stochastic coordinate descent is to pick at each step a direction e_j uniformly and to choose that e_j to be the direction of the descent at that step. More precisely, of I is drawn uniformly on $[d]$, then $\mathbb{E}[d\nabla_I f(x)e_I] = \nabla f(x)$. Therefore, the vector $d\nabla_I f(x)e_I$ that has only one nonzero coordinate is an unbiased estimate of the gradient $\nabla f(x)$. We can use this estimate to perform stochastic gradient descent.

Algorithm 3 Stochastic Coordinate Descent algorithm

```
**Input:** $x_1 \in \mathcal{C}$, positive sequence $\{\eta_s\}_{s\geq 1}$, independent random variables $I_1, \ldots, I_k$ uniform over $[d]$.
**for** $s = 1$ to $k-1$ **do**
    $y_{s+1} = x_s - \eta_s d\nabla_I f(x)e_I\,, \quad \tilde{g}_s \in \partial\ell(x_s, Z_s)$
    $x_{s+1} = \pi_{\mathcal{C}}(y_{s+1})$
**end for**
**return** $\bar{x}_k = \frac{1}{k}\sum_{s=1}^{k} x_s$
```

If we apply Stochastic Gradient Descent to this problem for $\eta = \frac{R}{L}\sqrt{\frac{2}{dk}}$, we directly obtain

$$
\mathbb{E}[f(\bar{x_k})] - f(x^*) \leq RL\sqrt{\frac{2d}{k}}
$$

We are in a trade-off situation where the updates are much easier to implement but where we need more steps to reach the same precision as the gradient descent alogrithm.

References

[Nem12] Arkadi Nemirovski, *On safe tractable approximations of chance constraints*, European J. Oper. Res. **219** (2012), no. 3, 707–718. MR 2898951 (2012m:90133)

[NS06] Arkadi Nemirovski and Alexander Shapiro, *Convex approximations of chance constrained programs*, SIAM J. Optim. **17** (2006), no. 4, 969–996. MR 2274500 (2007k:90077)