Exam Optimization under Uncertainty 4.2

December 2018

Problem 1 (10 Credits)

Let $J \in \mathcal{C}^2$ and consider the steepest descent algorithm :

$$\theta_{n+1} = \theta_n - \epsilon_n \nabla_\theta J(\theta_n). \tag{1}$$

Suppose that the gradient is bounded by some constant L, i.e., $\|\nabla_\theta J(\theta)\| \leq L$, for all $\theta \in \mathbb{R}^d$. Let $\epsilon_n = 1/(n+1)$ and show that

$$\|\theta_n\| \leq \|\theta_0\| + L(\ln(n+1)+1).$$

Hint: You may use that $\sum_{k=1}^{n} 1/k \leq \ln(n+1)+1$.

Answer Problem 1: By construction

$$\theta_n = \theta_0 - \sum_{k=0}^{n-1} \epsilon_k \nabla_\theta J(\theta_k).$$

Therefore,

$$\|\theta_n\| \leq \|\theta_0\| + \sum_{k=0}^{n-1} \|\epsilon_k \nabla_\theta J(\theta_k)\| \leq \|\theta_0\| + \sum_{k=0}^{n-1} \frac{1}{k+1} L = \|\theta_0\| + \sum_{k=1}^{n} \frac{1}{k} L = L(\ln(n+1)+1).$$

Problem 2 (total 10 Credits)

The gradient-field of a function $J(\theta)$ is shown in Figure 1. Apply a steepest descent algorithm for finding the minimum of $J(\theta)$.

(a). [5 Credits] Discuss with Figure 1 for the ODE

$$\frac{d}{dt} x(t) = -\nabla J(x(t))$$

the nature of point $(0,0)$ (stable, asymptotically stable, or unstable).

(b). [5 Credits] Judging from the range the figure, is this problem well-posed and is the vector-field coercive?

Answer Problem 2: (a)$(0,0)$ is a saddle point. Therefore, it is neither a stable nor is it asymptotically stable.

(b) The vector field points away from the origin along the diagonal (x,x) and $(-x,-x)$. Therefore, the problem is not well-posed (there seems to be no minimum) and the ODE is not coercive. The stable points of the ODE are no locations of a minimum.

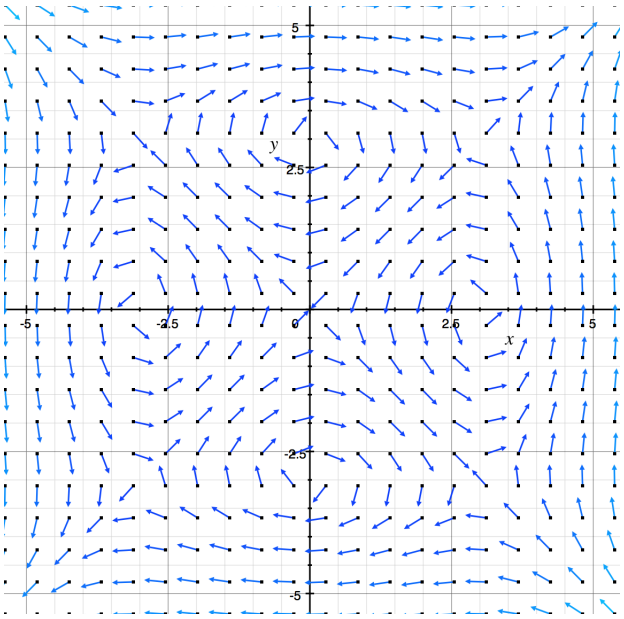

Figure 1: Gradient-field of $J(\theta)$

Problem 3 (total 20 Credits)

Consider the algorithm

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n,$$

for finding some optimal solution θ^*, for either $\epsilon_n \downarrow 0$ or $\epsilon_n = \epsilon$, for $n \in \mathbb{N}$. Suppose that evaluating Y_n requires one sample from an underlying process. Suppose your computational budget is sufficient to sample N samples from the underlying process.

(a). [5 Credits] Suppose you use your entire simulation budget to simulate one sample of θ_N. What conclusions can be drwan from this in case of decreasing ϵ and in case of fixed ϵ?

(b). [5 Credits] Suppose you split your simulation budget to produce k independent runs of the algorithm yielding $\theta_n(\omega_i)$, $1 \leq i \leq k$, for each of the runs, where $kn = N$. What conclusions can be drawn from this data in case of decreasing ϵ and in case of fixed ϵ?

(c). [10 Credits] For the given simulation budget N describe the best setup for your optimization algorithm that allows to produce a statistical justifiable assessment on the optimal solution θ^*.

Answer Problem 3: (a) For decreasing ϵ we obtain by θ_N an approximation of θ^* due to a.s. convergence of θ_n towards θ^*. For fixed ϵ we can draw no conclusion on θ^* due to the weak convergence.

(b) In both cases we can test for normality of θ_n and in case $\{\theta_i : 1 \leq i \leq k\}$ is approx. normal, we can construct a confidence interval for θ^*.

(c) There is a trade-off between number of independent replications and length of the individual runs. For decreasing ϵ, the lengths of the runs should be long enough so that the θ_n values are relatively stable. Independent replications are then used to produce confidence intervals for

θ^*. In the fixed ϵ case, we should take the runs long enough so the the average seems to become stable and use then use independent replications are then used to produce confidence intervals for θ^* (of course after checking for normality).

Problem 4 (total 30 Credits)

It takes an Erlang-$(3, 1/\theta)$-distributed time, denoted by $X(\theta)$, for a packet to traverse a communication channel. More formally, $X(\theta)$ has pdf

$$f(x, \theta) = \frac{1}{2\theta^3} x^2 e^{-x/\theta},$$

for $x \geq 0$, and $\mathbb{E}[X(\theta)] = 3\theta$. If the communication time exceeds a threshold value α, then the company has to pay a fine of c Euros per unit of excess time. The cost of operating the channel at "speed factor" θ is $1/\theta^2$. Let

$$J(\theta) = \mathbb{E}[c \max(X(\theta) - \alpha, 0)] + \frac{1}{\theta^2},$$

$\theta > 0$. Consider the problem

$$\min_{\theta > 0} J(\theta).$$

(a). [5 Credits] Argue that the problem is well-defined.

(b). [5 Credits] Take $G(\theta) = -dJ(\theta)/d\theta$ and argue that $G(\theta)$ is coercive for the optimization problem.

(c). [5 Credits] Compute the SF estimator for $dJ(\theta)/d\theta$ (you don't have to check unbiasedness).

(d). [5 Credits] Using the SF estimator from (a) provide a descent algorithm for finding the solution of the minimization problem.

(e). [5 Credits] You already argued that the problem is well-defined and the vector field is coercive. Letting $\epsilon_n = 1/(n+1)$, what properties have to be checked for establishing a.s. convergence of your algorithm to the location of the minimum?

(f). [5 Credits] Any descent direction will lead the algorithm Y_n to the solution. Provide an alternative coercive vector field such that the corresponding update $\tilde{Y}_n$ has less variance than your steepest descent algorithm in (d).

Answer Problem 4: (a) The part $\mathbb{E}[c \max(X(\theta) - \alpha)]$ is increasing in θ and minimization will push for small values of θ. The cost $1/\theta^2$ pushed for large values. Therefore, it is conceivable that the solution is at some point $\theta > 0$. We may replace the optimization problem by the equivalent problem:

$$\min J(\theta) \quad \text{s.t.} \quad \theta \in \Theta = \{\hat{\theta} : g(\hat{\theta}) = \delta - \hat{\theta} < 0\},$$

for some $\delta > 0$ small. By construction, the constraint $g(\theta) = \delta - \theta$ is not active. So the only remaining KKT point is the stationary point of $J(\theta)$.

(b) The mapping $J(\theta)$ seems to be smooth with continuous derivative. For $G(\theta) = -J'(\theta)$, the only stable point are the stationary points of $J(\theta)$. We show that the vector field stays bounded. Let

$$V(t) = J(x(t)) - J(\theta^*) \geq 0.$$

Differentiating with respect to t gives

$$\frac{d}{dt}V(t) = \frac{d}{dt}x(t)J'(x(t)) = -(J'(x(t))^2 < 0.$$

Therefore, $V(t)$ is bounded from below and monotone decreasing, and therefore has a limit. Hence,

$$V(0) \geq \sup_{t\geq 0} V(t) \geq \lim_{t\to\infty} V(t) \geq 0.$$

Therefore, $x(t)$ stays bounded along trajectories.

(c) The score function is obtained by

$$\frac{\partial}{\partial\theta}f(x,\theta) = \frac{\partial}{\partial\theta}\left(\frac{1}{2\theta^3}x^2e^{-x/\theta}\right) = \frac{x^2(x-3\theta)}{2\theta^5}e^{-x/\theta}.$$

Hence, we obtain for the Score Function

$$\mathrm{SF}(x,\theta) = \frac{x-3\theta}{\theta^2}.$$

(d) Let

$$Y_n = c\max(X(\theta_n)-\alpha,0)\mathrm{SF}(X(\theta_n),\theta_n) - \frac{2}{\theta_n^3}.$$

Noting that $X(\theta) = \theta X(1)$, we obtain

$$\begin{aligned}
Y_n &= c\max(X(\theta_n)-\alpha,0)\frac{X(\theta_n)-3\theta_n}{\theta_n^2} - \frac{2}{\theta_n^3} \\
&= c\max(\theta_n X(1)-\alpha,0)\frac{\theta_n X(1)-3\theta_n}{\theta_n^2} - \frac{2}{\theta_n^3} \\
&= c\max(\theta_n X(1)-\alpha,0)\frac{X(1)-3}{\theta_n} - \frac{2}{\theta_n^3} \\
&= \begin{cases} c\frac{(\theta_n X(1)-\alpha)(X(1)-3)}{\theta_n} - \frac{2}{\theta_n^3}, & \text{if } \theta_n X(1) > \alpha; \\ -\frac{2}{\theta_n^3}, & \text{otherwise.} \end{cases}
\end{aligned}$$

Then, the algorithm is

$$\theta_{n+1} = \theta_n - \epsilon_n Y_n.$$

(e) Given the estimator is unbiased. The key condition is

$$\sum_{n=1}^{\infty}\frac{1}{(n+1)^2}\mathbb{E}[(Y_n - J(\theta_n))^2|\mathcal{F}_{n-1}].$$

Since the variance is of order $1/\theta_n$, the variance control scheme has to be used, for example, for values of $\theta_n < 1$, to guarantee that the algorithm converges to the true minimum

(f) Typically IPA has less variance and taking $\tilde{Y}_n$ via an unbiased IPA estimator decreases the variance. Alternatively, you may take $\tilde{Y}_n$ as a batch means, i.e., average over a number of i.i.d. gradient samples, then the resulting algorithm has less variance.

In principle, the gradient estimator is just an implementation of the gradient field. An example of an alternative coercive vector field with less variance is, e.g., $\tilde{G}(\theta) = -\alpha J(\theta)$ for $0 < \alpha < 1$. Then the variance of Y_n is scaled by the factor α^2 and thus smaller.

Problem 5 (30 Credits)

Again consider $L(\theta) = \mathbb{E}\big[h\big(X(\theta)\big)\big]$, where

$$h\big(X(\theta)\big) = c \max\big\{X(\theta) - \alpha, 0\big\},$$

for some given $\alpha, c > 0$, and where $X(\theta)$ has an Erlang- $(3, 1/\theta)$-distribution for $\theta > 0$ (see Problem 4).

(a). [10 Credits] Recall your SF estimator of Problem 4(c) for being an unbiased estimator of $L'(\theta)$ (forget the $1/\theta^2$ in the objective function of Problem 4). To be unbiased a (nontrivial) condition is required concerning the derivative $\partial/\partial\theta f(x, \theta)$ of the pdf of $X(\theta)$. Formulate this condition. Then check that this condition holds in this problem.

(b). [5 Credits] Derive the MVD estimator $D^{\mathrm{MVD}}(\theta)$ of $L'(\theta)$.
Hint: it is a difference of two estimators involving Erlang- $(4, 1/\theta)$ and Erlang-$(3, 1/\theta)$ distributions.

(c). [10 Credits] Give the simulation algorithm for generating n replications of $D^{\mathrm{MVD}}(\theta)$, including how you generate from the Erlang-$(4, 1/\theta)$ and Erlang-$(3, 1/\theta)$ distributions. Make sure to exploit common random numbers as much as possible. From these n replications, give the sample average, the sample variance, and the standard error. Explain what the purpose is of these numbers.

(d). [5 Credits] Give the expression for a randomized MVD estimator $D^{\mathrm{MVDrand}}(\theta)$ that involves a single estimator (in stead of the difference in (b)). Show that $\mathbb{E}\big[D^{\mathrm{MVDrand}}(\theta)\big] = \mathbb{E}\big[D^{\mathrm{MVD}}(\theta)\big]$

Answer Problem 5:

(a). Suppose that we wish the derivative $L'(\cdot)$ in some $\theta_0 > 0$. Consider an open interval around θ_0: $0 < a < \theta_0 < b < \infty$. Then the condition is

$$\int_0^\infty |h(x)| \sup_{\theta\in(a,b)} \big|\frac{\partial}{\partial\theta} f(x, \theta)\big|\, dx < \infty.$$

First we compute the sup in any $x > 0$ (see Problem 4(c) for the derivative of the density function wrt θ):

$$\begin{aligned}
&\sup_{\theta\in(a,b)} \Big|\frac{\partial}{\partial\theta} f(x, \theta)\Big| \\
&= \sup_{\theta\in(a,b)} \Big|\frac{x^2(x - 3\theta)}{2\theta^5}\, e^{-x/\theta}\Big| \\
&= \sup_{\theta\in(a,b)} \Big|\big(\frac{x}{\theta} - 3\big)\frac{x^2}{2\theta^4}\, e^{-x/\theta}\Big| \\
&\le \Big(\sup_{\theta\in(a,b)} \big|\frac{x}{\theta} - 3\big|\Big)\frac{x^2}{2a^4}\, e^{-x/b} \\
&\le \big(\frac{x}{a} + 3\big)\frac{x^2}{2a^4}\, e^{-x/b}
\end{aligned}$$

Furthermore, $h(x) = c(x-\alpha)$ if $x > \alpha$, and otherwise $h(x) = 0$ Thus,

$$\int_0^\infty |h(x)| \sup_{\theta\in(a,b)} \big|\frac{\partial}{\partial\theta} f(x,\theta)\big|\, dx \le \int_\alpha^\infty c(x-\alpha)\,\big(\frac{x}{a}+3\big)\frac{x^2}{2a^4}\, e^{-x/b}\, dx.$$

The integrand is of the form $x^p e^{-x}$ which is integrable.

(b).

$$\frac{\partial}{\partial\theta} f(x,\theta) = \frac{x^2(x-3\theta)}{2\theta^5}\, e^{-x/\theta} = \frac{x^3}{2\theta^5}\, e^{-x/\theta} - \frac{3x^2}{2\theta^4}\, e^{-x/\theta}$$
$$= \frac{3}{\theta}\Big(\underbrace{\frac{x^3}{6\theta^4}\, e^{-x/\theta}}_{\text{Erlang-}\ (4,1/\theta)} - \underbrace{\frac{x^2}{2\theta^3}\, e^{-x/\theta}}_{\text{Erlang-}\ (3,1/\theta)} \Big).$$

Let $X^{(+)}(\theta)$ be the random variable with Erlang- $(4, 1/\theta)$ distribution, and $X^{(-)}(\theta)$ the random variable with Erlang- $(3, 1/\theta)$ distribution. Then

$$D^{\mathrm{MVD}}(\theta) = \frac{3}{\theta}\Big(h\big(X^{(+)}(\theta)\big) - h\big(X^{(-)}(\theta)\big)\Big).$$

(c). An Erlang- (k, λ) random variable is the sum of k independent Exponential-λ variables. Let $E_1(\theta), \ldots, E_4(\theta)$ be four i.i.d. random variables with Exponential-$1/\theta$ distribution. Then, we set $X^{(-)}(\theta) = \sum_{i=1}^3 E_i(\theta)$, and $X^{(+)}(\theta) = X^{(-)}(\theta) + E_4(\theta)$. Simulating an $E(\theta)$ is by applying the inverse transform: $E(\theta) = -\theta \ln(1-U)$, where U is uniform-$(0,1)$, obtained by a call of the random number generator. Thus,

Algorithm 1 MVD estimation

```
D ← zeros(n) {vector of zeros}
for k = 1 to n {generate n replications} do
    for i = 1 to 4 do
        U_i ~ U(0,1) {call RNG}
        E_i ← −θ ln(1 − U_i)
    end for
    X^(−) ← E_1 + E_2 + E_3
    h^(−) ← c max{X^(−) − α, 0}
    X^(+) ← X^(−) + E_4
    h^(+) ← c max{X^(+) − α, 0}
    D[k] ← (h^(+) − h^(−))/θ
end for
return vector D
```

The sample average

$$\overline{D_n}(\theta) = \frac{1}{n}\sum_{k=1}^n D_k(\theta)$$

is used for estimating $\mathbb{E}[D(\theta)] = L'(\theta)$. The sample variance

$$S^2(\theta) = \frac{1}{n-1}\sum_{k=1}^{n}\big(D_k(\theta) - \overline{D_n}(\theta)\big)^2$$

is used for estimating $\mathbb{V}ar(D(\theta))$. From this we estimate the standard error of $\overline{D_n}(\theta)$ by $\sqrt{S^2(\theta)/n}$. Call this σ, then $100(1-\alpha)\%$ confidence intervals are constructed by

$$\big(\overline{D_n}(\theta) - t_{n-1,1-\alpha/2}\sigma, \overline{D_n}(\theta) - t_{n-1,1-\alpha/2}\sigma\big),$$

where $t_{n-1,1-\alpha/2}$ is the $1-\alpha/2$-quantile of the student-t distribution with $n-1$ degrees of freedom, and where α is typically 0.05 or 0.1.

(d). In stead of computing both $X^{(+)}(\theta)$ and $X^{(-)}(\theta)$ in each iteration of the algorithm, we compute one of these chosen at random. Let $B \in \{0,1\}$ be a Bernoulli random variable with $\mathbb{P}(B=0) = \mathbb{P}(B=1) = 1/2$. Then

$$D^{\mathrm{MVDrand}}(\theta) = \frac{6}{\theta}\Big(h\big(X^{(+)}(\theta)\big)\mathbb{1}\{B=1\} - h\big(X^{(-)}(\theta)\big)\mathbb{1}\{B=0\}\Big).$$

Its expected value:

$$\begin{aligned}\mathbb{E}\big[D^{\mathrm{MVDrand}}(\theta)\big] &= \frac{6}{\theta}\Big(\mathbb{E}\big[h\big(X^{(+)}(\theta)\big)\big]\mathbb{P}(B=1) - \mathbb{E}\big[h\big(X^{(-)}(\theta)\big)\big]\mathbb{P}(B=0)\Big)\\ &= \frac{3}{\theta}\Big(\mathbb{E}\big[h\big(X^{(+)}(\theta)\big)\big] - \mathbb{E}\big[h\big(X^{(-)}(\theta)\big)\big]\Big) = \mathbb{E}\big[D^{\mathrm{MVD}}(\theta)\big]\end{aligned}$$

Bonus Question (total 10 Credits)

With the definitions and notation in place that have been introduced in Problem 3. Let

$$L(\theta) = \mathbb{E}[c\max(X(\theta) - \alpha, 0)],$$

$\theta > 0$. Consider the problem of finding θ^* such that

$$L(\theta^*) = \beta,$$

for some $\beta > 0$. Provide a descent algorithm and discuss sufficient condition for its convergence to θ^*.

Answer Bonus Question: Since $X(\theta)$ is monotone increasing in θ, $L(\theta)$ is monotone increasing. We let

$$G(\theta) = \beta - L(\theta).$$

Since $L(\theta)$ is monotone increasing, point θ^* is for the ODE

$$\frac{d}{dt}x(t) = \beta - L(x(t)) = G(x(t))$$

asymptotically stable. We thus consider

$$\theta_{n+1} = \theta_n + \epsilon_n(\beta - c\max(X(\theta_n) - \alpha, 0)).$$

We let $\epsilon_n = 1/(n+1)$ and

$$Y_n = \beta - c\max(X(\theta_n) - \alpha, 0)),$$

which is an unbiased estimator for $G(\theta_n)$, and it remains to ensure the variance condition. As usual,

$$V_n = \mathbb{E}\left[(\beta - c\max(X(\theta_n) - \alpha, 0) - G(\theta_n))^2 \,\middle|\, \mathcal{F}_{n-1}\right] = c^2\mathrm{Var}(\max(\mathrm{X}(\theta_\mathrm{n} - \alpha), 0).$$

For the variance conditon to hold we need

$$\sum_n \epsilon_n^2 V_n = c^2 \sum_n \frac{1}{(n+1)^2}\mathrm{Var}(\max(\mathrm{X}(\theta_\mathrm{n}) - \alpha, 0)$$

to be finite. The variance of $X(\theta_n)$ is not bounded on $(0, \infty)$ and we require the variance control scheme, i.e., replacing $\max(X(\theta_n), 0)$ by the sample average over k iid samples of $\max(X(\theta_n), 0)$.