Computer Intensive Methods in Mathematical Statistics

Johan Westerborn

Department of mathematics
KTH Royal Institute of Technology
johawes@kth.se

Lecture 4
Variance reduction for MC methods
30 March 2017

What do we need to know?

What do we need to master for having practical use of the MC method?

We agreed on that, for instance, the following questions should be answered:

1. How do we generate the needed input random variables?
2. How many computer experiments should we do? What can be said about the error?
3. Can we exploit problem structure to speed up the computation?

Outline

1 Introduction to variance reduction

2 Control variates

3 Antithetic sampling

Confidence interval from a simulation viewpoint

- Assume, as usual, that we estimate $\tau = \mathbb{E}(\phi(X))$ by means of MC, providing the level $1 - \alpha$ confidence interval

$$\mathsf{I}_\alpha = \left(\tau_N \pm \lambda_{\alpha/2} \frac{\sigma(\phi)}{\sqrt{N}} \right)$$

  for τ.

- Assume that we want to choose N large enough to assure that we estimate τ with an error less than a given $\varepsilon > 0$ on the specified level. This means that

$$\lambda_{\alpha/2} \frac{\sigma(\phi)}{\sqrt{N}} < \varepsilon \quad \Leftrightarrow \quad N > \lambda_{\alpha/2}^2 \frac{\sigma^2(\phi)}{\varepsilon^2}.$$

- Thus, the required MC sample size N (i.e., the required work) increases linearly with $\sigma^2(\phi)$.

Confidence interval from a simulation viewpoint

- Assume, as usual, that we estimate $\tau = \mathbb{E}(\phi(X))$ by means of MC, providing the level $1 - \alpha$ confidence interval

$$I_\alpha = \left(\tau_N \pm \lambda_{\alpha/2} \frac{\sigma(\phi)}{\sqrt{N}} \right)$$

  for τ.
- Assume that we want to choose N large enough to assure that we estimate τ with an error less than a given $\varepsilon > 0$ on the specified level. This means that

$$\lambda_{\alpha/2} \frac{\sigma(\phi)}{\sqrt{N}} < \varepsilon \quad \Leftrightarrow \quad N > \lambda_{\alpha/2}^2 \frac{\sigma^2(\phi)}{\varepsilon^2}.$$

- Thus, the required MC sample size N (i.e., the required work) increases linearly with $\sigma^2(\phi)$.

Confidence interval from a simulation viewpoint

- Assume, as usual, that we estimate $\tau = \mathbb{E}(\phi(X))$ by means of MC, providing the level $1 - \alpha$ confidence interval

$$I_\alpha = \left(\tau_N \pm \lambda_{\alpha/2} \frac{\sigma(\phi)}{\sqrt{N}} \right)$$

  for τ.
- Assume that we want to choose N large enough to assure that we estimate τ with an error less than a given $\varepsilon > 0$ on the specified level. This means that

$$\lambda_{\alpha/2} \frac{\sigma(\phi)}{\sqrt{N}} < \varepsilon \quad \Leftrightarrow \quad N > \lambda_{\alpha/2}^2 \frac{\sigma^2(\phi)}{\varepsilon^2}.$$

- Thus, the required MC sample size N (i.e., the required work) increases linearly with $\sigma^2(\phi)$.

Alternative representations of τ

- Thus, in general, a strategy to gain computational efficiency would thus be to find an alternative representation (ϕ', f') of τ, in the sense that

$$\tau = \mathbb{E}_f(\phi(X)) = \int_{\mathcal{X}} \phi(x) f(x)\, dx$$
$$= \int_{\mathcal{X}} \phi'(x) f'(x)\, dx = \mathbb{E}_{f'}(\phi'(X)),$$

  for which $\sigma^2_{f'}(\phi') < \sigma^2_f(\phi)$.

- Last time we saw that importance sampling (IS) was one way of achieving this: $f' \leftarrow g$ and $\phi' \leftarrow \phi\,\omega$.

Alternative representations of τ

- Thus, in general, a strategy to gain computational efficiency would thus be to find an alternative representation (ϕ', f') of τ, in the sense that

$$\tau = \mathbb{E}_f(\phi(X)) = \int_{\mathcal{X}} \phi(x) f(x)\, dx$$
$$= \int_{\mathcal{X}} \phi'(x) f'(x)\, dx = \mathbb{E}_{f'}(\phi'(X)),$$

  for which $\sigma^2_{f'}(\phi') < \sigma^2_f(\phi)$.
- Last time we saw that importance sampling (IS) was one way of achieving this: $f' \leftarrow g$ and $\phi' \leftarrow \phi\,\omega$.

Last time: Importance sampling

- The basis of importance sampling was to take an instrumental density g on X such that $g(x) = 0 \Rightarrow f(x) = 0$ and rewrite the integral as

$$\tau = \mathbb{E}_f\left(\phi(X)\right) = \int_{\mathsf{X}} \phi(x) f(x)\, dx = \int_{f(x)>0} \phi(x) f(x)\, dx$$
$$= \int_{g(x)>0} \phi(x) \frac{f(x)}{g(x)} g(x)\, dx = \mathbb{E}_g\left(\phi(X)\frac{f(X)}{g(X)}\right)$$
$$= \mathbb{E}_g\left(\phi(X)\omega(X)\right),$$

where

$$\omega : \{x \in \mathsf{X} : g(x) > 0\} \ni x \mapsto \frac{f(x)}{g(x)}$$

is the so-called importance weight function.

Last time: Importance sampling (cont.)

- Now estimate $\tau = \mathbb{E}_g(\phi(X)\omega(X))$ using standard MC:

  for $i = 1 \to N$ **do**
  draw $X^i \sim g$;
  end
  set $\tau_N^{\text{IS}} \leftarrow \sum_{i=1}^N \phi(X^i)\omega(X^i)/N$;
  return τ_N^{IS}

- The CLT provides immediately, as $N \to \infty$,

$$\sqrt{N}(\tau_N^{\text{IS}} - \tau) \xrightarrow{\text{d.}} \text{N}(0, \sigma_g^2(\phi\omega)),$$

  where $\sigma_g^2(\phi\omega) = \mathbb{V}_g(\phi(X)\omega(X))$ is estimated using `var`.

- Conclusion: Try to choose g so that the function $x \mapsto \phi(x)\omega(x)$ is close to constant in the support of g.

Last time: Self-normalized IS

- Often $f(x)$ is known only up to a normalizing constant $c > 0$, i.e. $f(x) = z(x)/c$, where we can evaluate $z(x) = cf(x)$ but not $f(x)$. We could then however show that τ can be rewritten as

$$\tau = \mathbb{E}_f(\phi(X)) = \ldots = \frac{\mathbb{E}_g(\phi(X)\omega(X))}{\mathbb{E}_g(\omega(X))},$$

where

$$\omega : \{x \in \mathsf{X} : g(x) > 0\} \ni x \mapsto \frac{z(x)}{g(x)}$$

is known and can be evaluated.

Last time: Self-normalized IS (cont.)

- Thus, having generated $X^1, \ldots, X^N$ from g we may estimate $\mathbb{E}_g(\phi(X)\omega(X))$ and $\mathbb{E}_g(\omega(X))$ using standard MC:

$$\tau = \frac{\mathbb{E}_g(\phi(X)\omega(X))}{\mathbb{E}_g(\omega(X))}$$

$$\approx \frac{\frac{1}{N}\sum_{i=1}^N \phi(X^i)\omega(X^i)}{\frac{1}{N}\sum_{\ell=1}^N \omega(X^\ell)} = \sum_{i=1}^N \underbrace{\frac{\omega(X^i)}{\sum_{\ell=1}^N \omega(X^\ell)}}_{\text{normalized weight}} \phi(X^i) = \tau_N^{\text{SNIS}}.$$

- As a by product we obtain the estimate

$$c = \mathbb{E}_g(\omega(X)) \approx \frac{1}{N}\sum_{\ell=1}^N \omega(X^\ell).$$

A CLT for self-normalized IS estimators

- One may establish the following CLT (see E2).

Theorem

Assume that $\sigma_g^2(\omega\phi) = \mathbb{V}_g(\omega(X)\phi(X)) < \infty$. *Then*

$$\sqrt{N}(\tau_N^{SNIS} - \tau) \xrightarrow{d.} \mathrm{N}(0, \sigma_g^2(\omega\{\phi - \tau\})/c^2),$$

where, as usual, $c = \mathbb{E}_g(\omega(X))$.

A CLT for self-normalized IS estimators (cont.)

- Here the asymptotic standard deviation

$$\sigma_g(\omega\{\phi - \tau\})/c$$

  is in general intractable.
- This quantity can however be estimated by
  1. letting, for each of the draws $X^i \sim g$,

$$Z_i = \omega(X^i)(\phi(X^i) - \tau_N^{\text{SNIS}}),$$

  2. applying `std` to the vector containing all the Z_is, and, finally,
  3. dividing the result by the MC estimate of the normalizing constant c.

Outline

1 Introduction to variance reduction

2 Control variates

3 Antithetic sampling

Control variates

- Assume that we have at hand another real-valued random variable Y, referred to as a control variate such that
  - (i) $\mathbb{E}(Y) = m$ is known and
  - (ii) Y can be simulated at the same complexity as $\phi(X)$.
- Then we may set, for some $\alpha \in \mathbb{R}$,

$$Z = \phi(X) + \alpha(Y - m),$$

so that

$$\mathbb{E}(Z) = \mathbb{E}(\phi(X) + \alpha(Y - m)) = \underbrace{\mathbb{E}(\phi(X))}_{=\tau} + \alpha \underbrace{(\mathbb{E}(Y) - m)}_{=0} = \tau.$$

Control variates

- Assume that we have at hand another real-valued random variable Y, referred to as a control variate such that
  - (i) $\mathbb{E}(Y) = m$ is known and
  - (ii) Y can be simulated at the same complexity as $\phi(X)$.
- Then we may set, for some $\alpha \in \mathbb{R}$,

$$Z = \phi(X) + \alpha(Y - m),$$

so that

$$\mathbb{E}(Z) = \mathbb{E}(\phi(X) + \alpha(Y - m)) = \underbrace{\mathbb{E}(\phi(X))}_{=\tau} + \alpha \underbrace{(\mathbb{E}(Y) - m)}_{=0} = \tau.$$

Control variates (cont.)

- In addition, if $\phi(X)$ and Y have covariance $\mathbb{C}(\phi(X), Y)$ it holds that

$$\begin{aligned}\mathbb{V}(Z) = \mathbb{V}(\phi(X) + \alpha Y) &= \mathbb{C}(\phi(X) + \alpha Y, \phi(X) + \alpha Y) \\ &= \mathbb{V}(\phi(X)) + 2\alpha\mathbb{C}(\phi(X), Y) + \alpha^2\mathbb{V}(Y).\end{aligned}$$

- Differentiating w.r.t. α and minimizing yields

$$0 = 2\mathbb{C}(\phi(X), Y) + 2\alpha\mathbb{V}(Y) \quad \Leftrightarrow \quad \alpha = \alpha^* = -\frac{\mathbb{C}(\phi(X), Y)}{\mathbb{V}(Y)},$$

which provides the optimal coefficient α^* in terms of variance.

Control variates (cont.)

- In addition, if $\phi(X)$ and Y have covariance $\mathbb{C}(\phi(X), Y)$ it holds that

$$\begin{aligned}\mathbb{V}(Z) = \mathbb{V}(\phi(X) + \alpha Y) &= \mathbb{C}(\phi(X) + \alpha Y, \phi(X) + \alpha Y) \\ &= \mathbb{V}(\phi(X)) + 2\alpha\mathbb{C}(\phi(X), Y) + \alpha^2\mathbb{V}(Y).\end{aligned}$$

- Differentiating w.r.t. α and minimizing yields

$$0 = 2\mathbb{C}(\phi(X), Y) + 2\alpha\mathbb{V}(Y) \quad \Leftrightarrow \quad \alpha = \alpha^* = -\frac{\mathbb{C}(\phi(X), Y)}{\mathbb{V}(Y)},$$

which provides the optimal coefficient α^* in terms of variance.

Control variates (cont.)

- Plugging α^* into the formula for $\mathbb{V}(Z)$ gives

$$\mathbb{V}(Z) = \mathbb{V}(\phi(X)) + 2\alpha^*\mathbb{C}(\phi(X), Y) + (\alpha^*)^2\mathbb{V}(Y) = \ldots$$
$$= \mathbb{V}(\phi(X))\left(1 - \frac{\mathbb{C}(\phi(X), Y)^2}{\mathbb{V}(\phi(X))\mathbb{V}(Y)}\right) = \mathbb{V}(\phi(X))\{1-\rho(\phi(X), Y)^2\},$$

  where

$$\rho(\phi(X), Y) \stackrel{\text{def}}{=} \frac{\mathbb{C}(\phi(X), Y)}{\sqrt{\mathbb{V}(\phi(X))}\sqrt{\mathbb{V}(Y)}}$$

  is the correlation between $\phi(X)$ and Y.

- Consequently, we can expect large variance reduction if $|\rho(\phi(X), Y)|$ is close to 1.

Control variates (cont.)

- Plugging α^* into the formula for $\mathbb{V}(Z)$ gives

$$\mathbb{V}(Z) = \mathbb{V}(\phi(X)) + 2\alpha^*\mathbb{C}(\phi(X), Y) + (\alpha^*)^2\mathbb{V}(Y) = \ldots$$
$$= \mathbb{V}(\phi(X))\left(1 - \frac{\mathbb{C}(\phi(X), Y)^2}{\mathbb{V}(\phi(X))\mathbb{V}(Y)}\right) = \mathbb{V}(\phi(X))\{1-\rho(\phi(X), Y)^2\},$$

  where

$$\rho(\phi(X), Y) \stackrel{\text{def}}{=} \frac{\mathbb{C}(\phi(X), Y)}{\sqrt{\mathbb{V}(\phi(X))}\sqrt{\mathbb{V}(Y)}}$$

  is the correlation between $\phi(X)$ and Y.

- Consequently, we can expect large variance reduction if $|\rho(\phi(X), Y)|$ is close to 1.

Example: another tricky integral

- As an example, estimate

$$\tau = \int_0^{\pi/2} \exp(\cos^2(x))\,dx = \int_0^{\pi/2} \underbrace{\frac{\pi}{2}\exp(\cos^2(x))}_{=\phi(x)} \underbrace{\frac{2}{\pi}}_{=f(x)}\,dx$$
$$= \mathbb{E}_f(\phi(X))$$

using

$$Z = \phi(X) + \alpha^*(Y - m),$$

where $Y = \cos^2(X)$ is a control variate with

$$m = \mathbb{E}(Y) = \int_0^{\pi/2} \cos^2(x)\frac{2}{\pi}\,dx = \{\text{integration by parts}\} = \frac{1}{2}$$

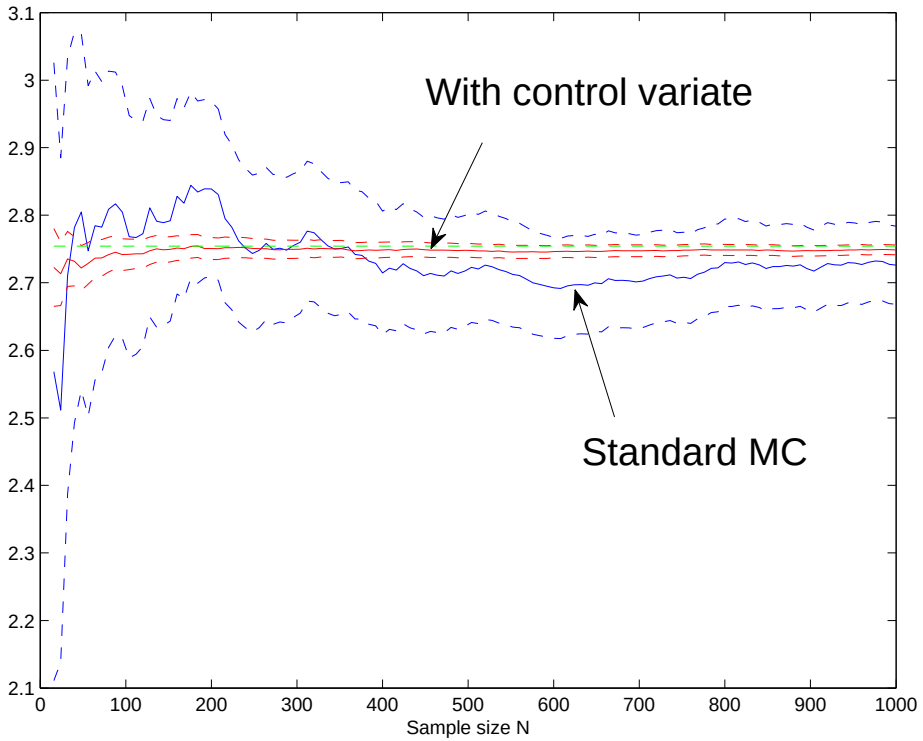
and $\hat{\alpha}^*$ is an estimate of the optimal coefficient.

Example: another tricky integral (cont.)

- In MATLAB:

```
cos2 = @(x) cos(x).^2;
phi = @(x) (pi/2)*exp(cos2(x));
X = (pi/2)*rand(1,N);
tau = mean(phi(X));
Y = cos2(X);
m = 1/2;
alpha = - cov([phi(X)' Y'])./var(Y); % appr. optimal alpha
Z = phi(X) + alpha(1,2)*(Y - m);
tau_CV = mean(Z);
```

Example: another tricky integral (cont.)



Outline

1 Introduction to variance reduction

2 Control variates

3 Antithetic sampling

Antithetic sampling

Again, assume that we wish to estimate $\tau = \mathbb{E}_f(\phi(X))$ by means of MC. For simplicity, introduce the short-hand notation $V \stackrel{\text{def}}{=} \phi(X)$, so that $\tau = \mathbb{E}(V)$.

Now, assume we can generate another variable V' such that

(i) $\mathbb{E}(V') = \tau$,

(ii) $\mathbb{V}(V') = \mathbb{V}(V)$ $(= \sigma^2(\phi))$,

(iii) V' can be simulated at the same complexity as V.

Antithetic sampling (cont.)

- Then, letting

$$W \stackrel{\text{def}}{=} \frac{V + V'}{2},$$

  it holds that $\mathbb{E}(W) = \tau$.

- Moreover,

$$\mathbb{V}(W) = \mathbb{V}\left(\frac{V + V'}{2}\right) = \frac{1}{4}\left(\mathbb{V}(V) + 2\mathbb{C}(V, V') + \mathbb{V}(V')\right)$$
$$= \frac{1}{2}\left(\mathbb{V}(V) + \mathbb{C}(V, V')\right).$$

Antithetic sampling (cont.)

- Now, note that each W is twice as costly to generate as each V.
- Thus, for a fixed computation budget we can choose between generating $2N$ Vs or N Ws.
- In terms confidence bounds it is better to use the Ws if

$$\lambda_{\alpha/2}\frac{\mathbb{D}(W)}{\sqrt{N}} < \lambda_{\alpha/2}\frac{\mathbb{D}(V)}{\sqrt{2N}} \Leftrightarrow 2\mathbb{V}(W) < \mathbb{V}(V)$$

$$\Leftrightarrow \mathbb{V}(V) + \mathbb{C}(V, V') < \mathbb{V}(V)$$

$$\Leftrightarrow \mathbb{C}(V, V') < 0.$$

- Thus, if we can find V' such that the antithetic variables V and V' are negatively correlated, then we will gain computational work.

Antithetic sampling (cont.)

- Now, note that each W is twice as costly to generate as each V.
- Thus, for a fixed computation budget we can choose between generating $2N$ Vs or N Ws.
- In terms confidence bounds it is better to use the Ws if

$$\lambda_{\alpha/2}\frac{\mathbb{D}(W)}{\sqrt{N}} < \lambda_{\alpha/2}\frac{\mathbb{D}(V)}{\sqrt{2N}} \Leftrightarrow 2\mathbb{V}(W) < \mathbb{V}(V)$$

$$\Leftrightarrow \mathbb{V}(V) + \mathbb{C}(V, V') < \mathbb{V}(V)$$

$$\Leftrightarrow \mathbb{C}(V, V') < 0.$$

- Thus, if we can find V' such that the antithetic variables V and V' are negatively correlated, then we will gain computational work.

Antithetic sampling (cont.)

- Now, note that each W is twice as costly to generate as each V.
- Thus, for a fixed computation budget we can choose between generating $2N$ Vs or N Ws.
- In terms confidence bounds it is better to use the Ws if

$$\lambda_{\alpha/2}\frac{\mathbb{D}(W)}{\sqrt{N}} < \lambda_{\alpha/2}\frac{\mathbb{D}(V)}{\sqrt{2N}} \Leftrightarrow 2\mathbb{V}(W) < \mathbb{V}(V)$$

$$\Leftrightarrow \mathbb{V}(V) + \mathbb{C}(V, V') < \mathbb{V}(V)$$

$$\Leftrightarrow \mathbb{C}(V, V') < 0.$$

- Thus, if we can find V' such that the antithetic variables V and V' are negatively correlated, then we will gain computational work.

Antithetic sampling (cont.)

- For this purpose, the following theorem can be very useful.

Theorem

Let U be a random variable and let $\varphi : \mathbb{R} \to \mathbb{R}$ be a monotone function. Moreover, assume that there exists a non-increasing transform $T : \mathbb{R} \to \mathbb{R}$ such that $U \overset{d.}{=} T(U)$. Then $V = \varphi(U)$ and $V' = \varphi(T(U))$ are identically distributed and

$$\mathbb{C}(V, V') = \mathbb{C}(\varphi(U), \varphi(T(U))) \leq 0.$$

Example: a tricky integral (reconsidered)

- We estimate again

$$\tau = \int_0^{\pi/2} \exp(\cos^2(x))\,dx = \int_0^{\pi/2} \underbrace{\frac{\pi}{2}\exp(\cos^2(x))}_{=\phi(x)} \underbrace{\frac{2}{\pi}}_{=f(x)}\,dx$$
$$= \mathbb{E}(\phi(X)) = \mathbb{E}(\varphi(U)),$$

where $\begin{cases} U = \cos^2(X), \\ \varphi(u) = \frac{\pi}{2}\exp(u). \end{cases}$

- Now letting $T(u) = 1 - u$ yields

$$T(U) = 1 - \cos^2(X) = \sin^2(X)$$
$$= \cos^2\left(\frac{\pi}{2} - X\right) \overset{\text{d.}}{=} \cos^2(X) = U.$$

Example: a tricky integral (reconsidered)

- We estimate again

$$\tau = \int_0^{\pi/2} \exp(\cos^2(x))\,dx = \int_0^{\pi/2} \underbrace{\frac{\pi}{2}\exp(\cos^2(x))}_{=\phi(x)} \underbrace{\frac{2}{\pi}}_{=f(x)}\,dx$$
$$= \mathbb{E}(\phi(X)) = \mathbb{E}(\varphi(U)),$$

where $\begin{cases} U = \cos^2(X), \\ \varphi(u) = \frac{\pi}{2}\exp(u). \end{cases}$

- Now letting $T(u) = 1 - u$ yields

$$T(U) = 1 - \cos^2(X) = \sin^2(X)$$
$$= \cos^2\left(\frac{\pi}{2} - X\right) \overset{\text{d.}}{=} \cos^2(X) = U.$$

Example: a tricky integral (reconsidered)

- Since in addition $T(u) = 1 - u$ is non-increasing and $\varphi(u) = \frac{\pi}{2}\exp(u)$ monotone, the theorem above applies. Thus,

$$\mathbb{C}\left(\frac{\pi}{2}\exp(\cos^2(X)), \frac{\pi}{2}\exp(\underbrace{1-\cos^2(X)}_{=\sin^2(X)})\right) \leq 0,$$

and we may apply antithetic sampling with

$$\begin{cases} V = \frac{\pi}{2}\exp(\cos^2(X)), \\ V' = \frac{\pi}{2}\exp(\sin^2(X)), \\ W = \frac{V+V'}{2}. \end{cases}$$

Example: a tricky integral (reconsidered) (cont.)

- In Matlab:

```
cos2 = @(x) cos(x).^2;
phi = @(x) (pi/2)*exp(cos2(x));
X = (pi/2)*rand(1,N);
tau = mean(phi(X));
XX = (pi/2)*rand(1,N/2); % only half the sample size
V_1 = (pi/2)*exp(cos2(XX));
V_2 = (pi/2)*exp(1 - cos2(XX));
W = (V_1 + V_2)/2;
tau_AS = mean(W);
UB = tau + norminv(0.975)*std(phi(X))./sqrt(N);
LB = tau - norminv(0.975)*std(phi(X))./sqrt(N);
UB_AS = tau_AS + norminv(0.975)*std(W)./sqrt(N/2);
LB_AS = tau_AS - norminv(0.975)*std(W)./sqrt(N/2);
```

Example: a tricky integral (reconsidered) (cont.)