Mathematics for New Technologies in Finance

Solution sheet 6

Exercise 6.1 (Portfolio optimization) See notebook 1.

Solution 6.1 See notebook 1.

Exercise 6.2 (Optimal portfolio allocation) Consider the Merton's optmization problem outlined in lecture notebook 5. Define the set of admissible controls as $\mathcal{A} = \mathbb{H}_2$, that is to say the set of $\mathbb{F}$-adapted processes such that $\mathbb{E}\left[\int_0^T |\alpha_s|^2 \mathrm{d}s\right] < +\infty$. Fix also the parameters $\{r, \mu, \sigma\} \in \mathbb{R}^3_+$. Define, for $\gamma \in (0,1)$, the the CRRA (Constant Relative Risk Aversion) function $u = \frac{(\cdot)^\gamma - 1}{\gamma}$. The limit case $\gamma = 0$ corresponds to $u = \log$. We aim to solve the optimization problem

$$\tilde{V}(t,x) := \sup_{\alpha \in \mathcal{A}} \mathbb{E}[u(X_T^{\alpha,t,x})],$$

where the process $X^{\alpha,t,x}$ starts at time t with initial value x.

(a) Write down explicitly the dynamics of the wealth process X^α.

(b) Define

$$V(t,x) := \sup_{\alpha \in \mathcal{A}} \mathbb{E}[(X_T^{\alpha,t,x})^\gamma],$$

and notice that $\tilde{V}(t,x) = \frac{V(t,x) - 1}{\gamma}$. Verify that the value function V solves the following PDE, known as dynamic programming equation:

$$\begin{aligned} \frac{\partial w}{\partial t}(t,x) &= -\sup_{a \in \mathbb{R}} \left[(r + (\mu - r)a) x \frac{\partial w}{\partial x}(t,x) + \frac{1}{2}\sigma^2 a^2 x^2 \frac{\partial^2 w}{\partial x^2}(t,x) \right] \\ w(T,x) &= x^\gamma \end{aligned} \tag{1}$$

(c) Using the ansatz $V(t,x) = x^\gamma h(t)$, reduce 1 to an ODE and solve it explicitly. Deduce the optimal Merton's ratio $\alpha^\star$, the explicit expression of V and the one of $\tilde{V}$.

Solution 6.2

(a)

$$\begin{aligned} dX_t^\alpha &= \alpha_t X_t \frac{dS_t}{S_t} + (1 - \alpha_t) X_t \frac{dS_t^0}{S_t^0} \\ &= [r + (\mu - r)\alpha_t] X_t dt + \sigma \alpha_t X_t dW_t. \end{aligned} \tag{2}$$

(b) This is a classical problem in stochastic control, associating a deterministic equation (PDE) to the value function of a control problem, given the payoff $\mathbb{E}[(X_T^{\alpha,t,x})^\gamma]$ and the dynamics 2. A derivation of the Hamilton-Jacobi-Bellman equation, which is essentially a consequence of the dynamic programming principle, can be found in Chapters 2 and 4 of [1]. We will not delve too much in details here, but point out that the PDE, in case $\mathcal{A}$ is reduced to a singleton, boils down to the celebrated Feynman–Kac formula.

(c) Using the suggested ansatz, we write $v(t,x) = x^\gamma h(t)$. Equation 1 is reduced to

$$\begin{aligned} 0 &= h' + \gamma h \sup_{a \in \mathbb{R}} \left\{ r + (\mu - r)a + \frac{1}{2}(\gamma - 1)\sigma^2 a^2 \right\}, \\ h(T) &= 1. \end{aligned}$$

It is easy to compute the optimal $a^\star$, which is

$$a^\star = \frac{\mu - r}{(1-\gamma)\sigma}.$$

The ODE is then reduced to

$$\begin{aligned} 0 &= h' + \gamma h \left[r + \frac{1}{2} \frac{(\mu - r)^2}{(1-\gamma)\sigma^2} \right], \\ h(T) &= 1. \end{aligned}$$

This leads to the unique candidate:

$$h^\star(t) := e^{k(T-t)} \quad \text{with} \quad k := \gamma \left[r + \frac{1}{2} \frac{(\mu - r)^2}{(1-\gamma)\sigma^2} \right].$$

We have now constructed a solution v to equation 1, but a priori it may not coincide with V. Indeed, we only know that V is also a solution to 1. By a verification argument (see Chapter 4 of [1]), we can indeed conclude that $V(t,x) = v(t,x) = h^\star(t)x$. We recall $\tilde{V}(t,x) = \frac{V(t,x)-1}{\gamma}$. Finally, the Merton's ratio does not depend on time and it is given as follows:

$$\alpha_t^\star = a^\star = \frac{\mu - r}{(1-\gamma)\sigma}, \quad \forall t \in [0,T].$$

Exercise 6.3 (Convergence of norms) Consider a measure space $(\Omega, \mathcal{M}, \sigma)$ and a measurable function $f : \Omega \to \mathbb{R}$, with $f \in \mathcal{L}^p(\Omega) \cap \mathcal{L}^\infty(\Omega)$ for some $0 < p < +\infty$.

(a) Prove that $\lim_{q \to +\infty} \|f\|_q = \|f\|_\infty$.

(b) If we do not assume explicitly that $f \in \mathcal{L}^p(\Omega)$, the statement may not hold anymore. Under which other assumption the statement is true, without directly assuming $f \in \mathcal{L}^p(\Omega)$?

Solution 6.3

(a) If $\|f\|_\infty = 0$, the proposition holds trivially. Otherwise, fix $\epsilon > 0$ and $S_\epsilon := \{x : |f(x)| \geq \|f\|_\infty - \epsilon\}$, for $\epsilon < \|f\|_\infty$. This set has positive measure (by definition of $\|\cdot\|_\infty$) and, by the fact that $f \in \mathcal{L}^p(\Omega)$, its measure is finite. Indeed,

$$\infty > \|f\|_p \geqslant \left(\int_{S_\epsilon} (\|f\|_\infty - \epsilon)^p \, d\mu \right)^{1/p} = (\|f\|_\infty - \epsilon) \, \mu \left(S_\epsilon \right)^{1/p}.$$

We now use the same estimate, but for a generic $q > 0$. We send first $q \to +\infty$, leading to $\liminf_{q \to +\infty} \|f\|_q \geq (\|f\|_\infty - \epsilon)$, and finally, sending $\epsilon \searrow 0$, we get

$$\liminf_{q \to +\infty} \|f\|_q \geq \|f\|_\infty.$$

For the converse inequality, we recall that $|f(x)| \leq \|f\|_\infty$ for almost every x. Then, for $q > p$,

$$\|f\|_q \leqslant \left(\int_X |f(x)|^{q-p} |f(x)|^p d\mu \right)^{1/q} \leqslant \|f\|_\infty^{\frac{q-p}{q}} \|f\|_p^{p/q}.$$

Notice that, by assumption, all the elements on the right-hand side are finite. The required inequality is obtained once we send $q \to +\infty$.

(b) Take $\Omega = \mathbb{R}$, $\mathcal{M} = \mathcal{B}(\mathbb{R})$, and $\mu = Leb(\mathbb{R})$. Then, $f \equiv 1$ gives the counterexample. One may assume that $\mu(\Omega) < +\infty$; this assumption, together with $f \in \mathcal{L}^\infty(\Omega)$, implies $f \in \mathcal{L}^p(\Omega)$ for any $p \leq +\infty$.

Exercise 6.4 (Bayesian optimization)

(a) Recall the definition of prior, likelihood, posterior, and evidence distributions in Bayesian statistics.

(b) Consider linear model on $\mathbb{R} : Y \sim \theta X + Z, \theta \sim \mathcal{N}(0,1), Z \sim \mathcal{N}(0,1)$ and θ independent with X. Compute $p_\theta(y \mid x)$ and $p(\theta \mid x, y)$. Prove that maximizing the posterior $p(\theta \mid x, y)$ is exactly doing Ridge regression (fix λ here).

(c) Consider Lasso regression, what is the prior under Bayesian perspective? Please calculate the posterior under this prior.

(d) Would you expect a sparser weight or denser weight using Lasso regression instead of Ridge regression.

Solution 6.4

(a) Assume that we want to infer the distribution of X, which depends on a unknown parameter θ. We call $\pi = \pi(\theta)$ the prior distribution on the parameter θ, which corresponds to the initial guess on the parameter, before inferring on it based on the realization of X. Notice we assume that all the distributions we work with admit a density. The likelihood $f_\theta = f_\theta(x)$ corresponds to the distribution of X, given θ. In particular, we notice that the joint distribution of (X, θ) can be express as $f_\theta(x)\pi(\theta)\mathrm{d}x\mathrm{d}\theta$. Finally, the evidence corresponds to the marginal distribution $f = f(x)$ of X, while the posterior $\pi(\theta|x)$ to the distribution of θ given the realization $X = x$. By Bayes' theorem, we have

$$\pi(\theta|x) = \frac{f_\theta(x)\pi(\theta)}{f(x)} = \frac{f_\theta(x)\pi(\theta)}{\int_\Theta f_\theta(x)\pi(\theta)\mathrm{d}\theta}.$$

(b) See the proof here.

(c) Suppose we have data points $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$. where $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$. The likelihood for the data is

$$\mathcal{L}(Y \mid X, \beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2\right)$$

We select the prior for β to follow the Laplace distribution (also known as double-exponential distribution) with a zero mean and common scale parameter $b : \pi(\beta) = (1/2b)\exp(-|\beta|/b)$. Multiplying the prior distribution with the likelihood we get the posterior distribution

$$\mathcal{L}(Y \mid X, \beta)\pi(\beta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2\right)\left[\frac{1}{2b}\exp\left(-\frac{|\beta|}{b}\right)\right].$$

Notice that this proof corresponds exactly to the one of previous point, once we change the prior distribution accordingly.

(d) We recall first, given the linear model $Y = \theta X + Z$, where $Z \sim \mathcal{N}(0, \mathbb{I})$, Ridge and Lasso regression correspond to minimizing, respectively,

$$\theta_R^\star := \arg\min_{\theta\in\Theta} \|Y - X\theta\|_2^2 + \lambda_R \|\theta\|_2;$$
$$\theta_L^\star := \arg\min_{\theta\in\Theta} \|Y - X\theta\|_2^2 + \lambda_L \|\theta\|_1,$$

where λ_R and λ_L are scaling parameters. If now $\Theta = \mathbb{R}^d$, it can be proved that there exists two parameters K_R and K_L such that

$$\theta_R^\star = \arg\min_{\|\theta\|_2 \le K_R} \|Y - X\theta\|_2^2;$$
$$\theta_L^\star = \arg\min_{\|\theta\|_1 \le K_L} \|Y - X\theta\|_2^2.$$

The shape of balls in $\|\cdot\|_1$-norm explains now why the Lasso regularization induces sparsity in the weight θ (in the sense that more components of θ are null), compared to the Ridge penalization.

Exercise 6.5 (Stochastic gradient descent)

(a) Assume that we aim to find the θ^* to maximize the posterior:

$$p(\theta|x_1,\cdots,x_n) = \frac{p(\theta)\prod_{i=1}^n p(x_i|\theta)}{p(x_1,\cdots,x_n)}$$

with stochastic gradient descent method in practice. In each step, do we calculate $\nabla p(\theta|x_1,\cdots,x_n)$? do we calculate $\nabla \log p(\theta|x_1,\cdots,x_n)$? do we calculate $\nabla \log p(\theta)$ or $\nabla \log p(x_i|\theta)$?

(b) If $p(x_1,\cdots,x_n)$ has no closed formula, does it cause a trouble when we do stochastic gradient descent?

(c) Construct a stochastic differential equation with invariant measure to be the posterior distribution $p(\theta|x_1,\cdots,x_n)$.

Solution 6.5

(a) In stochastic gradient descent, one updates the parameter subject to

$$\theta^{[t+1]} = \theta^{[t]} - \delta\nabla_\theta \mathcal{L}_X(\theta^{[t]}).$$

Hence, the optimization boils down to the study of the posterior $p(\theta|x_1,\cdots,x_n)$. The usage of the log does not change the optimization, but allows to decouple the terms. In particular,

$$\sup_{\theta\in\Theta} p(\theta|x_1,\cdots,x_n) = \sup_{\theta\in\Theta}\log p(\theta|x_1,\cdots,x_n) = \sup_{\theta\in\Theta}\left[\log p(\theta) + \sum_{i=1}^n \log p(x_i|\theta) - \log p(x_1,\cdots,x_n)\right]. \tag{3}$$

In particular, we only have to calculate $\nabla \log p(\theta)$ and $\nabla \log p(x_i|\theta)$.

(b) It is clear from 3 that $\log p(x_1,\cdots,x_n)$ is only a scaling term, which does not affect the optimization process.

(c) See lecture notebook 3.

References

[1] Nizar Touzi. *Optimal Stochastic Control, Stochastic Target Problems, and Backward SDE.* Springer, 2013.