Preliminary Notes on the Interplay Between Estimation and Optimization Problems

Thomas B. Schön
Division of Automatic Control
Linköping University
SE–58183 Linköping, Sweden.
E-mail: schon@isy.liu.se

January 28, 2009

Contents

Chapter 1

Introduction

Let x denote the variable we seek to estimate and y denote the variables that we have measured. Then we have

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \propto p(y|x)p(x) \tag{1.1}$$

where $p(y|x)$ is referred to as the likelihood, $p(x)$ is referred to as the prior density and $p(x|y)$ is referred to as the posterior density. Depending on how we choose to model the unknown variable x we end up with different results. Let us start by assuming that the unknown variable x is a stochastic variable distributed according to a certain probability density function (PDF) $p(x)$, the *prior* PDF. This is known as the Bayesian approach to estimation. When the measurements y are available we ask for the most probable value of the unknown variable x that can be explained by the measurements. This implies that the problem to be solved is

$$\hat{x}^{\text{MAP}} = \arg\max_x p(x|y) \tag{1.2}$$

which according to (1.1) is equivalent to

$$\hat{x}^{\text{MAP}} = \arg\max_x p(y|x)p(x) \tag{1.3}$$

The estimate $\hat{x}^{\text{MAP}}$ is referred to as the *Maximum A Posteriori* (MAP) estimate. Within the *Maximum Likelihood* (ML) framework, introduced by Fisher (1912), it is instead assumed that the unknown variable x is an unknown deterministic variable. Another way of stating this assumption is to say that the prior $p(x)$ is completely uninformative, implying that we have no prior knowledge of the behaviour of the unknown variable. The corresponding estimator is given simply by maximizing the likelihood function,

$$\hat{x}^{\text{ML}} = \arg\max_x p(y|x) \tag{1.4}$$

The intuition is that we are looking for the variables x that best explains the measurements y, in the sense that they make the measurements as likely as possible.

From the above discussion it is clear that an estimation problem sooner or later is transformed into an optimization problem of some form. We will here stress the use of convex optimization problems, which have the nice property that any local optimum is also the global optimum. Even though most problems are non-convex, they are often solved by suitable convex approximations, as we shall see later. Another advantage of casting the estimation problem as a convex optimization problem is that it is straightforward to add constraints to the problem. The theory on convex optimization is by now rather mature and there is general purpose software[1] available for solving the resulting problems. In this way prior information about the state can be utilized, e.g., that the state is always positive or that the components of the state should sum to one, which is the case if the state is a vector of probabilities. Constraints of this type cannot be straightforwardly included in the standard Kalman filter. However, if we can (which we can, as will be shown later) formulate the Kalman filter as an optimization problem, it is straightforward to add arbitrary convex constraints to this problem and still guaranteeing a global optimum to be found.

The main message in convex optimization is that we should *not* differ between linear and nonlinear optimization problems, but instead between convex and non-convex problems. The class of convex problems is much larger than that covered by linear problems,

A convex optimization problem is defined as

$$\begin{aligned} \min_{x} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \qquad i = 0, \ldots, m, \\ & a_j^T x = b_j, \qquad j = 0, \ldots, n, \end{aligned} \tag{1.5}$$

where the functions $f_0, \ldots, f_m$ are convex and the equality constraints are linear. For a thorough introduction to convex optimization, see Boyd and Vandenberghe (2004).

It is also worth stressing that it is straightforward to include other variables to be estimated, such as, e.g., missing data into the optimization problem. Besides including them in the variables to be estimated, there is probably also a need to provide some assumptions regarding how they behave, which are typically implemented as constraints.

[1] A useful and efficient software is YALMIP, developed by Löfberg (2006). It provides direct access to several of the standard numerical solvers for optimization problems, using a powerful MATLAB interface.

Chapter 2

State Estimation as an Optimization Problem

During the lecture on nonlinear state estimation we derived a conceptual solution to the nonlinear sequential state estimation problem, see Schön (2008) for details. Here we will take a different route, rather than insisting an directly finding a sequential solution we will formulate an optimization problem containing all the state variables. We can then of course study sequential solutions to this problem and recover the solutions derived differently before. However, more importantly, we can straightforwardly consider new problems, i.e., constrained state estimation. Furthermore, this way of modelling problems typically reveals a lot of the structure inherent in the problem. Once this structure is well understood we can try to find efficient approximations, rather than the other way around. This is particularly true for the SLAM problem, which is very briefly discussed in Chapter 4.

2.1 Problem Formulation

We are considering a nonlinear state-space model with additive noise according to

$$x_{t+1} = f(x_t) + w_t, \tag{2.1a}$$
$$y_t = h(x_t) + e_t, \tag{2.1b}$$

where the noise w_t and e_t are i.i.d. Furthermore, the initial state x_1 is random, with $x_1 \sim \mathcal{N}(\bar{x}_1, \bar{P}_1)$. Note that everything can be straightforwardly generalized to time-varying models, possibly containing a known control input u_t as well. However, in the interest of a simple notation we consider the time-invariant case (2.1). The *goal* is to estimate the states $x_{1:t}$ given the measurements $y_{1:t}$. We will do this by considering the MAP estimate,

$$\hat{x}_{1:t} = \arg\max_{x_{1:t}} p(x_{1:t}|y_{1:t}), \tag{2.2}$$

where

$$p(x_{1:t}|y_{1:t}) = \frac{p(y_{1:t}|x_{1:t})p(x_{1:t})}{p(y_{1:t})} \propto p(y_{1:t}|x_{1:t})p(x_{1:t}) \tag{2.3}$$

Since the measurement noise e_t is i.i.d. we have

$$p(y_{1:t}|x_{1:t}) = \prod_{i=1}^{t} p(y_i|x_i). \tag{2.4}$$

Furthermore, the likelihood $p(y_i|x_i)$ is given by

$$p(y_i|x_i) = p_{e_i}(y_i - h(x_i)). \tag{2.5}$$

since the measurement noise e_t enters additively in (2.1). In order to find a manageable expression for $p(x_{1:t})$ we start by making use of the product rule according to

$$p(x_{1:t}) = p(x_t, x_{1:t-1}) = p(x_t|x_{1:t-1})p(x_{1:t-1}), \tag{2.6}$$

which by the Markov property is reduced to

$$p(x_{1:t}) = p(x_t|x_{t-1})p(x_{1:t-1}). \tag{2.7}$$

Repeated use of (2.6) and (2.7) results in

$$p(x_{1:t}) = p(x_1)\prod_{i=2}^{t} p(x_i|x_{i-1}), \tag{2.8}$$

where $p(x_1)$ denote the prior on the state at time $t = 1$. Similarly to (2.5), the additive process noise w_t implies

$$p(x_i|x_{i-1}) = p_{w_{i-1}}(x_i - f(x_{i-1})). \tag{2.9}$$

Assembling (2.3) – (2.5) and (2.7) results in

$$p(x_{1:t}|y_{1:t}) \propto p(x_1)\prod_{i=2}^{t} p(x_i|x_{i-1})\prod_{i=1}^{t} p(y_i|x_i) \tag{2.10a}$$

$$= p_{x_1}(x_1 - \bar{x}_1)\prod_{i=2}^{t} p_{w_{i-1}}(x_i - f(x_{i-1}))\prod_{i=1}^{t} p_{e_i}(y_i - h(x_i)). \tag{2.10b}$$

To sum up, we have now arrived at an expression for the MAP estimator (2.2) that we can work with

$$\hat{x}_{1:t} = \arg\max_{x_{1:t}} p_{x_1}(x_1 - \bar{x}_1)\prod_{i=2}^{t} p_{w_{i-1}}(x_i - f(x_{i-1}))\prod_{i=1}^{t} p_{e_i}(y_i - h(x_i)) \tag{2.11}$$

The estimates arising from (2.11) are in the form

$$\hat{x}_{i|t}, \qquad i = 1, \ldots, t. \tag{2.12}$$

In other words, the estimates are *smoothed*, save for the last $\hat{x}_{t|t}$ which is of course the filtered estimate.

2.2 Relevant Special Cases

2.2.1 Linear Gaussian State-Space Model

A very commonly used special case of (2.1) is the linear (f and h are linear functions) state-space model, subject to Gaussian (w_t and e_t are Gaussian) noise,

$$x_{t+1} = Ax_t + w_t, \qquad w_t \sim \mathcal{N}(0, Q), \tag{2.13a}$$
$$y_t = Cx_t + e_t, \qquad e_t \sim \mathcal{N}(0, R). \tag{2.13b}$$

This implies that the densities involved in (2.11) are given by

$$p_{x_1}(x_1 - \bar{x}_1) \propto e^{-\frac{1}{2}\|x_1 - \bar{x}_1\|^2_{\bar{P}_1^{-1}}}, \tag{2.14a}$$
$$p_{w_{i-1}}(x_i - f(x_{i-1})) \propto e^{-\frac{1}{2}\|x_i - Ax_{i-1}\|^2_{Q^{-1}}}, \tag{2.14b}$$
$$p_{e_i}(y_i - h(x_i)) \propto e^{-\frac{1}{2}\|y_i - Cx_i\|^2_{R^{-1}}}, \tag{2.14c}$$

Using the fact that the logarithm is a monotonic function and the fact that maximizing a function is equivalent to minimizing the negated function we have

$$\max_x p(x) \qquad \Leftrightarrow \qquad \min_x -\log p(x) \tag{2.15}$$

The resulting MAP estimation problem is now given by

$$\hat{x}_{1:t} = \arg\min_{x_{1:t}} \|x_1 - \bar{x}_1\|^2_{\bar{P}_1^{-1}} + \sum_{i=2}^{t} \|x_i - Ax_{i-1}\|^2_{Q^{-1}} + \sum_{i=1}^{t} \|y_i - Cx_i\|^2_{R^{-1}} \tag{2.16}$$

This is a convex optimization problem, more specifically it is a *quadratic program* (QP). The theory on how to handle least-squares problems of this type is well established, see e.g., Björck (1996) and the many references therein. Here it is worth noting that we can prove that the sequential solution to (2.16) is given by the Kalman filter. Hence, one interpretation of the Kalman filter is as a *sequential solution to a weighted least squares problem.* Furthermore, we are free to add any convex constraints we like to (2.16) and still have a guarantee of finding the global optima. This is further formalized in Section 2.3. The size of the problem cast in (2.16) grows as the time increases. This can be handled very much in the same way as the problem size is controlled within a Model Predictive Controller (MPC), i.e., we simply bound the number of variables allowed in the optimization problem. More specifically this corresponds to solving the optimization problem over a sliding window, commonly referred to as moving horizon estimation (MHE), which is further formalized in Section 2.4. There is in fact an interesting duality between the control and the estimation problem, which unfortunately is out of scope for this discussion, but the interested reader is referred to for example (Goodwin et al., 2005; Kailath et al., 2000).

2.2.2 Nonlinear Gaussian State-Space Model

Let us now consider the nonlinear state-space, subject to additive Gaussian noise, i.e., (2.1) where $w_t \sim \mathcal{N}(0, Q)$ and $e_t \sim \mathcal{N}(0, R)$. Hence, the correspond-

ing MAP problem is the following nonlinear least squares problem

$$\hat{x}_{1:t} = \underset{x_{1:t}}{\arg\min} \ \|x_1 - \bar{x}_1\|^2_{\bar{P}_1^{-1}} + \sum_{i=2}^{t} \|x_i - f(x_{i-1})\|^2_{Q^{-1}} + \sum_{i=1}^{t} \|y_i - h(x_i)\|^2_{R^{-1}} \tag{2.17}$$

Methods commonly used in solving (2.17), such as Gauss-Newton and Levenberg-Marquart, rely on linear approximations of the involved nonlinear functions. A solid account on how to solve nonlinear least square problems is provided by (Nocedal and Wright, 2006; Dennis and Schnabel, 1983).

Using a first order Taylor expansion of the nonlinear functions f and h results in a linear approximation according to,

$$f(x_t) \approx f(x_t^*) + \underbrace{\frac{\partial f(x_t)}{\partial x_t}\Big|_{x_t=x_t^*}}_{F_t}(x_t - x_t^*) = f(x_t^*) + F_t\Delta_{x_t}, \tag{2.18a}$$

$$h(x_t) \approx h(x_t^*) + \underbrace{\frac{\partial h(x_t)}{\partial x_t}\Big|_{x_t=x_t^*}}_{H_t}(x_t - x_t^*) = h(x_t^*) + H_t\Delta_{x_t}, \tag{2.18b}$$

where x_t^* denotes the linearization point and $\Delta_{x_t} = x_t - x_t^*$ denotes the deviation from the linearization point. The linear approximation of the dynamics is given by

$$\begin{aligned} \|x_t - f(x_{t-1})\|^2_{Q^{-1}} &\approx \left\|x_t - f(x_{t-1}^*) - F_{t-1}\Delta_{x_{t-1}}\right\|^2_{Q^{-1}} \\ &= \left\|x_t - x_t^* + x_t^* - f(x_{t-1}^*) - F_{t-1}\Delta_{x_{t-1}}\right\|^2_{Q^{-1}} \\ &= \left\|\Delta_{x_t} - F_{t-1}\Delta_{x_{t-1}} - a_t\right\|^2_{Q^{-1}}, \end{aligned} \tag{2.19}$$

where

$$a_t = f(x_{t-1}^*) - x_t^*. \tag{2.20}$$

Similarly for the measurement relations we have

$$\|y_t - h(x_t)\|^2_{R^{-1}} \approx \|H_t\Delta_{x_t} - c_t\|^2_{R^{-1}}, \tag{2.21}$$

where

$$c_t = h(x_t^*) - y_t. \tag{2.22}$$

To sum up we have

$$\widehat{\Delta}_{x_{1:t}} = \underset{\Delta_{x_{1:t}}}{\arg\max} \ V(\Delta_{x_{1:t}}) \tag{2.23a}$$

where

$$\begin{aligned} V(\Delta_{x_{1:t}}) = \|\Delta_{x_1}\|^2_{\bar{P}_1} &+ \sum_{i=2}^{t} \|\Delta_{x_i} - F_{i-1}\Delta_{x_{i-1}} - a_i\|^2_{Q^{-1}} \\ &+ \sum_{i=1}^{t} \|H_i\Delta_{x_i} - c_i\|^2_{R^{-1}} \end{aligned} \tag{2.23b}$$

This problem will quickly become high-dimensional. Luckily it is a rather sparse problem, due to the inherent structure present in the problem. In order to make this obvious we will rewrite (2.23a), in the form

$$\widehat{\Delta} = \arg\max_{\Delta} V(\Delta) \tag{2.24}$$

$$V(\Delta) = \|A\Delta - b\|_2^2 \tag{2.25}$$

where $A \in \mathbb{R}^{t(n_x+n_y)\times tn_x}$ is a sparse matrix and $b \in \mathbb{R}^{t(n_x+n_y)}$. This is a problem that has been thoroughly studied within the linear algebra community, see e.g., Golub and Van Loan (1996); Björck (1996).

Let the square root of the symmetric positive semidefinite matrix Q be defined as $Q \triangleq Q^{T/2}Q^{1/2}$, allowing us to rewrite the weighted norm $\|x\|_{Q^{-1}}^2$ as an unweighted norm $\|x\|_2^2$ according to

$$\begin{aligned}\|x\|_{Q^{-1}}^2 &= x^T Q^{-1} x = x^T (Q^{T/2}Q^{1/2})^{-1} x = x^T Q^{-1/2} Q^{-T/2} x \\ &= (Q^{-T/2}x)^T (Q^{-T/2}x) = \|Q^{-T/2}x\|_2^2\end{aligned} \tag{2.26}$$

This can be used in order to normalize the expression in (2.23b), i.e., rewrite it as an unweighted norm according to

$$\begin{aligned}V(\Delta_{x_{1:t}}) &= \|\bar{P}_1^{-T/2}\Delta_{x_1}\|_2^2 + \sum_{i=2}^{t} \left\| Q^{-T/2}(\Delta_{x_i} - F_{i-1}\Delta_{x_{i-1}} - a_i) \right\|_2^2 \\ &\quad + \sum_{i=1}^{t} \left\| R^{-T/2}(H_i\Delta_{x_i} - c_i) \right\|_2^2 \\ &= \left\| \tilde{P}_1 \Delta_{x_1} \right\|_2^2 + \sum_{i=2}^{t} \left\| \begin{pmatrix} \tilde{F}_{i-1} & \tilde{Q} \end{pmatrix} \begin{pmatrix} \Delta_{x_{i-1}} \\ \Delta_{x_i} \end{pmatrix} - \tilde{a}_i \right\|_2^2 \\ &\quad + \sum_{i=1}^{t} \left\| \tilde{H}_i \Delta_{x_i} - \tilde{c}_i \right\|_2^2,\end{aligned} \tag{2.27}$$

where

$$\tilde{P}_1 = \bar{P}_1^{-T/2} \in \mathbb{R}^{n_x \times n_x}, \qquad \tilde{a}_i = Q^{-T/2} a_i \in \mathbb{R}^{n_x} \tag{2.28a}$$

$$\tilde{Q} = Q^{-T/2} \in \mathbb{R}^{n_x \times n_x}, \qquad \tilde{c}_i = R^{-T/2} c_i \in \mathbb{R}^{n_y} \tag{2.28b}$$

$$\tilde{F}_i = -Q^{-T/2} F_i \in \mathbb{R}^{n_x \times n_x}, \qquad \tilde{H}_i = R^{-T/2} H_i \in \mathbb{R}^{n_y \times n_x}. \tag{2.28c}$$

Let us now define

$$\Delta = \begin{pmatrix} \Delta_{x_1}^T & \Delta_{x_2}^T & \dots & \Delta_{x_t}^T \end{pmatrix}^T \in \mathbb{R}^{tn_x} \tag{2.29}$$

$$A = \begin{pmatrix} \tilde{P}_1 & 0 & 0 & \dots & 0 & 0 \\ \widetilde{F}_1 & \widetilde{Q} & 0 & \dots & 0 & 0 \\ 0 & \widetilde{F}_2 & \widetilde{Q} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \widetilde{F}_{t-1} & \widetilde{Q} \\ \widetilde{H}_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \widetilde{H}_2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & \widetilde{H}_t \end{pmatrix}, \qquad b = \begin{pmatrix} 0 \\ \tilde{c}_1 \\ \tilde{c}_2 \\ \vdots \\ \tilde{c}_t \\ \tilde{a}_1 \\ \tilde{a}_2 \\ \vdots \\ \tilde{a}_t \end{pmatrix} \tag{2.30}$$

allowing us to rewrite (2.27) according to (2.24).

2.3 Constrained State Estimation

An interesting class of problems is given in 1 below.

Problem 1 (Convex optimization filtering)

Assume that the densities $p_{x_0}(x_0)$, $p_{w_i}(w_i)$, and $p_{e_i}(e_i)$ are log-concave[1]*. In the presence of constraints in terms of a linear dynamic model 2.13, the MAP-estimate is the solution $\hat{x}_t = x_t$ to the following problem*

$$\begin{aligned} \max_{X_t} \quad & \log(p_{x_0}(x_0 - \bar{x}_0)) + \sum_{i=0}^{t-1} \log(p_{w_i}(w_i)) + \sum_{i=0}^{t} \log(p_{e_i}(e_i)) \\ \text{s.t.} \quad & x_{i+1} = A_i x_i + w_i, \qquad i = 0, \dots, t-1, \\ & y_i = C_i x_i + e_i, \qquad i = 0, \dots, t. \end{aligned}$$

It is straightforward to add any convex constraints to this formulation, and the resulting problem can be solved using standard software.

The reason why this is an interesting class of problems is that they are convex and we can straightforwardly add any convex constraints to the problem and sole for the global optimum. An examples of this is given in Schön et al. (2003).

2.4 Moving Horizon State Estimation

The main concern with the formulation of the estimation problem given so far is that the number of variables increase linearly in time, which is of course not acceptable in many applications. Hence, we have to find a way of bounding the number of variables. If possible we can of course derive a sequential solution, but in many cases this is impossible. Another way of bounding the number of variables in the optimization problem is to use *moving horizon estimation* (MHE) (Maciejowski, 2002; Goodwin et al., 2005), defined in Problem 2. This is the same idea underpinning model predictive control (MPC), i.e., the state is estimated using a fixed size, moving window of data.

[1] A function $f : \mathbb{R}^n \to \mathbb{R}$ is *log-concave* if $f(x) > 0$ for all x in the domain of f and $\log(f)$ is a concave function (Boyd and Vandenberghe, 2004).

Problem 2 (Moving Horizon Estimation (MHE))

Assume that the densities $p_{w_i}(w_i)$ and $p_{e_i}(e_i)$ are log-concave. In the presence of constraints in terms of a linear dynamic model, the MHE-estimate is the solution $\hat{x}_t = x_t$ to the following problem

$$\begin{aligned} \min_{x_{t-L:t}} \quad & F(x_{t-L}) + \sum_{i=i-L}^{t-1} \|x_i - f(x_{i-1})\|^2_{Q^{-1}} + \sum_{i=t-L+1}^{t} \|y_i - h(x_i)\|^2_{R^{-1}} \\ \text{s.t.} \quad & f_i(x_{t-L:t}) \leq 0, \qquad i = 0, \ldots, m, \\ & a_j^T x_{t_L:t} = 0, \qquad j = 0, \ldots, n, \end{aligned}$$

where $F(x_{t-L})$ contains information about the past. Additional constraints are included via the convex functions $f_1, \ldots, f_m$ convex and the linear equality constraints.

The problem is now reduced to solving a possible nonlinear and possibly constrained least squares problem, with a fixed number of variables. This problem has to be solved each time a new measurement arrives. However, it is important to understand that the approach using MHE is, in general, sub-optimal, since the influence of the past measurements is not necessarily taken care of correctly in $F(x_{t-L})$.

Several useful entry points into the literature on moving horizon estimation for nonlinear systems are given in Rao et al. (2001); Rao (2000); Goodwin et al. (2005).

Chapter 3

Linear Regression

Consider the classical linear regression problem,

$$y_t = \varphi_t^T \theta + e_t, \qquad t = 1, \ldots, N, \tag{3.1}$$

where $y_t \in \mathbb{R}$ denotes the measurement, $\varphi_t \in \mathbb{R}^n$ denotes the regressor, $\theta \in \mathbb{R}^d$ denotes the unknown, $e_t \in \mathbb{R}$ denotes the measurement noise and N denotes the number of samples. We assume that we have observed N measurements $y_{1:N} = \{y_t\}_{t=1}^N$ and that the regressors $\varphi_{1:t}$ are available. Furthermore, we assume that the measurement noise e_t is independent and identically distributed (iid) and Gaussian, $e_t \sim \mathcal{N}(0, \sigma^2), \sigma > 0$. The N equations in (3.1) can just as well be stacked on top of each other and written according to

$$Y = \Phi^T \theta + E, \tag{3.2}$$

where

$$Y = \begin{pmatrix} y_1 & y_2 & \ldots & y_N \end{pmatrix}^T \in \mathbb{R}^N \tag{3.3a}$$

$$\Phi = \begin{pmatrix} \varphi_1 & \varphi_2 & \ldots & \varphi_N \end{pmatrix} \in \mathbb{R}^{n \times d} \tag{3.3b}$$

$$E = \begin{pmatrix} e_1 & e_2 & \ldots & e_N \end{pmatrix}^T \in \mathbb{R}^N \tag{3.3c}$$

The problem of estimating the unknown θ has been extensively studied over the years. We will in this chapter discuss the linear regression problem using a few common model assumptions and estimators. As we will see, all problems end up in the form

$$\arg\min_\theta \ \|Y - \Phi^T \theta\|_2^2 + \lambda \|\theta\|_p^p \tag{3.4}$$

for different choices of λ and p.

3.1 Maximum Likelihood Estimation

The Maximum Likelihood estimate $\hat{\theta}^{\mathrm{ML}}$ is give by

$$\hat{\theta}^{\mathrm{ML}} = \arg\max_\theta \ p(Y|\theta) \tag{3.5}$$

According to the assumptions above the measurement noise is Gaussian and independent, implying that the likelihood function $p(Y|\theta)$ is given by

$$p(Y|\theta) = p_E(Y - \Phi^T\theta) = \mathcal{N}(Y\,;\,\Phi^T\theta, R) \propto e^{-\frac{1}{2}\|Y-\Phi^T\theta\|^2_{R^{-1}}}, \tag{3.6}$$

where $R = \sigma^2 I_N$. Inserting (3.6) into (3.5) results in

$$\hat{\theta}^{\mathrm{ML}} = \arg\max_{\theta}\ e^{-\frac{1}{2}\|Y-\Phi^T\theta\|^2_{R^{-1}}} \tag{3.7}$$

Since the logarithm is a monotonic function we can just as well minimize the log-likelihood function,

$$\hat{\theta}^{\mathrm{ML}} = \arg\min_{\theta}\ \left\|Y - \Phi^T\theta\right\|^2_{R^{-1}} \tag{3.8}$$

We can rewrite this problem as an unweighted problem by noting that

$$\|Y - \Phi^T\theta\|^2_{R^{-1}} = \left\|R^{-T/2}(Y - \Phi^T\theta)\right\|^2_2 = \frac{1}{\sigma^2}\left\|Y - \Phi^T\theta\right\|^2_2 \tag{3.9}$$

where we have used the fact that $R = \sigma^2 I_N$. Noting that σ is not part of the optimization variable and inserting (3.9) into (3.8) finally gives us the following ordinary least squares problem

$$\hat{\theta}^{\mathrm{ML}} = \arg\min_{\theta}\ \left\|Y - \Phi^T\theta\right\|^2_2 \tag{3.10}$$

which allows for a closed-form solution

$$\hat{\theta}^{\mathrm{ML}} = (\Phi\Phi^T)^{-1}\Phi Y. \tag{3.11}$$

3.2 Maximum a Posteriori Estimation

Let us now study the MAP estimate (1.3) for the linear regression problem (3.2)

$$\hat{\theta}^{\mathrm{MAP}} = \arg\max_{\theta}\ p(Y|\theta)p(\theta) \tag{3.12}$$

Let us write down a problem that is equivalent to (3.12) and directly useful in derivations to follow,

$$\hat{\theta}^{\mathrm{MAP}} = \arg\min_{\theta}\ -\log p(Y|\theta) - \log p(\theta) \tag{3.13}$$

The first term in (3.13) is according to (3.9) given by

$$-\log p(Y|\theta) \propto \frac{1}{\sigma^2}\|Y - \Phi^2\theta\|^2_2. \tag{3.14}$$

In Section 3.2.1 we consider the problem resulting from a Gaussian prior and in Section 3.2.2 we consider a Laplacian prior.

3.2.1 Gaussian Prior

Consider the situation where the individual parameters $\{\theta_i\}_{i=1}^d$ are independent, each having a Gaussian prior, according to

$$\theta_i \sim \mathcal{N}(0, \sigma^2/\lambda), \qquad i = 1, 2, \ldots, d. \tag{3.15}$$

The independece assumtions implies that

$$-\log p(\theta) = \prod_{i=1}^{d} p(\theta_i) = \sum_{i=1}^{d} -\log p(\theta_i), \tag{3.16}$$

which using the Gaussian assumption (3.15) can be written as

$$-\log p(\theta) = \sum_{i=1}^{d} \frac{\lambda}{\sigma^2}\theta_i^2 = \frac{\lambda}{\sigma^2}\|\theta\|_2^2 \tag{3.17}$$

Using (3.14) and (3.17), the resulting problem is given by

$$\hat{\theta}^{\text{MAP}} = \arg\max_{\theta} \ \left\|Y - \Phi^T\theta\right\|_2^2 + \lambda\|\theta\|_2^2 \tag{3.18}$$

Similarly to (3.10) this problem allows for a closed-form solution and it is straightforward to show that it is given by

$$\hat{\theta} = (\Phi\Phi^T + \lambda I_N)^{-1}\Phi Y. \tag{3.19}$$

The estimator given in (3.18) is within the field of statistics referred to as *ridge regression* and in other fields is travels under names such as *regularized least squares* and *Tikhonov regularization*. This estimator is often used for its ability to deal with outliers and for ill-posed problems it is used to ensure that an inverse exists.

For intuition it is worth noting that as the variance tends to infinity (λ tends to 0) (3.19) is reduced to the ML problem (3.11).

3.2.2 Laplacian Prior

Let us now consider the situation where the individual parameters $\{\theta_i\}_{i=1}^d$ are independent, each and having a Laplacian prior,

$$\theta_i \sim \mathcal{L}(0, 2\sigma^2/\lambda), \qquad i = 1, 2, \ldots, d. \tag{3.20}$$

This implies that

$$-\log p(\theta) = -\sum_{i=1}^{d} \frac{\lambda}{2\sigma^2} e^{-\frac{\lambda}{\sigma^2}|\theta_i|} \propto \frac{\lambda}{\sigma^2}\|\theta\|_1 \tag{3.21}$$

Using (3.14) and (3.21) results in

$$\hat{\theta}^{\text{MAP}} = \arg\max_{\theta} \ \left\|Y - \Phi^T\theta\right\|_2^2 + \lambda\|\theta\|_1 \tag{3.22}$$

which does not allow for a closed-form solution. However, the problem (3.22) is convex, guaranteeing that any minima is the global minima. There exists good methods and software for solving this type of problem, see e.g., Kim et al. (2007). The estimator (3.22) is referred to as the Least Absolute Shrinking and Selection Operator (LASSO), introduced by Tibshirani (1996). Typically the LASSO solution results in a sparse estimate $\hat{\theta}$. Recently this sparseness property of the l_1 norm has spurred a lot of interest under names such as compressed sensing, compressed sampling and l_1 magic.

3.3 Summary and Useful Convex Extensions

In Table 3.1 we summarize the discussion in the previous sections. The problems

Table 3.1: *Cost functions and the corresponding probabilistic assumptions.*

Cost function	Estimator	Prior
$\|Y - \Phi^T\theta\|_2^2$	ML	-
$\|Y - \Phi^T\theta\|_2^2 + \lambda\|\theta\|_2^2$	MAP	$\mathcal{N}(0, \sigma^2/\lambda)$
$\|Y - \Phi^T\theta\|_2^2 + \lambda\|\theta\|_1$	MAP	$\mathcal{L}(0, 2\sigma^2/\lambda)$

derived above are all in the form

$$\min_x \quad f_0(x) \tag{3.23}$$

where depending on the assumptions made, the cost function $f_0(x)$ is the cost function used in (3.10),(3.18) or (3.22). Similarly to what was done for the state estimation problem, if there is additional knowledge available about the problem, this should of course be used in forming the estimates. The best is of course if the constraints can be expressed in terms of a convex set, since then the results problem will be convex,

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \qquad i = 0, \ldots, m, \\ & a_j^T x = b_j, \qquad j = 0, \ldots, n, \end{aligned} \tag{3.24}$$

where the functions $f_0, \ldots, f_m$ are convex and the equality constraints are linear.

Non-convex constraints can of course be used as well, resulting in a much harder problem.

Chapter 4

Least Squares Formulations of SLAM and VO

This chapter is very sketchy and very short, but some of the details not showed during the lecture are present. It will be expanded as more of this has been applied. Currently this section is heavily inspired by Dellaert and Kaess (2006).

In order to mathematically describe the data association, let us introduce a data association variable

$$c_{tj} \in \{1, 2, \ldots, N+1\}, \tag{4.1}$$

where N is the total number of landmarks. The data association variable encodes the relation between the measurement y_{tj} and landmark l_k, where $k = c_{tj}$. In other words,

$$c_{tj} = i \leq N \tag{4.2}$$

implies that the j^{th} measurement at time t is associated to landmark i.

The measurement equation is provided by the camera model according to

$$y_{tj} = h(x_t, l_{c_{tj}}) + e_{tj}, \qquad e_{tj} \sim \mathcal{N}(0, R) \tag{4.3}$$

where $h(x_t, l_{c_{tj}}) = \mathcal{P}(L^w_{c_{tj}})$, derived during lecture 2. Note that we can of course use different sensors.

Similar to what was done in Section 2.2.2 we will now formulate a nonlinear least squares problem, that we solve using standard methods, involving linearization. According to (4.3) we have the following likelihood function

$$p(y_{tj}|x_t, l_{c_{tj}}) \propto e^{-\frac{1}{2}\|y_{tj} - h(x_t, l_{c_{tj}})\|^2_{R^{-1}}} \tag{4.4}$$

The full SLAM problem now corresponds to the following nonlinear least squares problem

$$\underset{x_{1:t}, l_{1:N}}{\arg\min} \; \|x_1 - \bar{x}_1\|^2_{\bar{P}_1^{-1}} + \sum_{i=2}^{t} \|x_i - f(x_{i-1})\|^2_{Q^{-1}} + \sum_{i=1}^{t} \sum_{j=1}^{M_i} \|y_{ij} - h(x_i, l_{c_{ij}})\|^2_{R^{-1}} \tag{4.5}$$

which we can approximate according to

$$\underset{x_{1:t},l_{1:N}}{\arg\min}\ V(x_{1:t}, l_{1:N}) \tag{4.6}$$

where

$$\begin{aligned} V(x_{1:t}, l_{1:N}) = \|x_1 - \bar{x}_1\|^2_{\bar{P}_1^{-1}} + \sum_{i=2}^{t} \|\Delta_{x_i} - F_{i-1}\Delta_{x_{i-1}} - a_i\|^2_{Q^{-1}} \\ + \sum_{i=1}^{t}\sum_{j=1}^{M_i} \|H_i^{c_{ij}}\Delta_{x_i} + L_i^{c_{ij}}\Delta_{l_{c_{ij}}} - d_i\|^2_{R^{-1}} \end{aligned} \tag{4.7}$$

Appendix A

Useful Facts from Probability Theory

In this appendix we list some facts from probability theory that are important to understand in order to read the main text. For a detailed treatment of probability theory we refer to one of the many textbooks on the subject.

Theorem 1 (Bayes' theorem) *Given two random variables a and b the conditional probability density function $p(a|b)$ is given by*

$$p(a|b) = \frac{p(b|a)p(a)}{p(b)}. \tag{A.1}$$

Definition 1 (Markov property) A discrete-time stochastic process $\{x_t\}$ is said to possess the Markov property if

$$p(x_{t+1}|x_1, \ldots, x_t) = p(x_{t+1}|x_t). \tag{A.2}$$

Definition 2 (Multivariable Normal Density) *A random variable x with $\mathrm{E}\{x\} = \mu_x$ and $\mathrm{Cov}\{x\} = \Sigma_x$, such that $\det \Sigma_x > 0$ is $\mathcal{N}(\mu_x, \Sigma_x)$ if and only if the probability density function for x is*

$$p(x) = \frac{1}{(2\pi)^{n_x/2}\sqrt{\det \Sigma_x}} e^{-\frac{1}{2}(x-\mu_x)^T \Sigma_x^{-1}(x-\mu_x)} \tag{A.3}$$

In order to have a practical notation for stating that a probability density function is normal with a certain mean value and covariance we will define the following

$$\mathcal{N}(x\,;\,\mu_x, \Sigma_x) = \frac{1}{(2\pi)^{n_x/2}\sqrt{\det \Sigma_x}} e^{-\frac{1}{2}(x-\mu_x)^T \Sigma_x^{-1}(x-\mu_x)}. \tag{A.4}$$

This notation will allow us to write $\mathcal{N}(x\,;\,\mu_x, \Sigma_x)$ rather than the entire expression in (A.4), which is of course convenient.

Definition 3 (Laplace Density) *A random variable x is said to be Laplacian distributed if the density function for x is*

$$p(x) = \frac{1}{2\sigma} e^{-\frac{|x-\mu|}{\sigma}}, \tag{A.5}$$

The Laplace density is also referred to as the double exponential density. Similarly to what we did for the Gaussian density above, we will use the following notation for the Laplacian density,

$$\mathcal{L}(x\,;\,\mu,\sigma) = \frac{1}{2\sigma} e^{-\frac{|x-\mu|}{\sigma}}, \tag{A.6}$$

where $\mu \in \mathbb{R}$ is the location parameter and $\sigma > 0$ is the scale parameter.

Bibliography

Björck, Å. (1996). *Numerical methods for least squares problems.* SIAM, Philadelphia, PA, USA.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization.* Cambridge University Press, Cambridge, UK.

Dellaert, F. and Kaess, M. (2006). Square root SAM: Simultaneous localization and mapping via square root information smoothing. *The International Journal of Robotics Research*, 25(12):1181–1204.

Dennis, J. E. and Schnabel, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice Hall.

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41:155–160.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations.* John Hopkins University Press, Baltimore, third edition.

Goodwin, G. C., Seron, M. M., and De Doná, J. A. (2005). *Constrained Control and Estimation An Optimisation Approach.* Communications and Control Engineering. Springer, London, UK.

Kailath, T., Sayed, A. H., and Hassibi, B. (2000). *Linear Estimation.* Information and System Sciences Series. Prentice Hall, Upper Saddle River, NJ, USA.

Kim, S.-J., Koh, K., Lustig, M., Boyd, S., and Gorinevsky, D. (2007). An interior-point method fort large-scale l_1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):606–617.

Löfberg, J. (2006). YALMIP: Software for solving convex (and nonconvex) optimization problems. In *Proceedings of the American Control Conference (ACC)*, Minneapolis, MN, USA.

Maciejowski, J. M. (2002). *Predictive Control with Constraints.* Prentice Hall.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization.* Springer Series in Operations Research. Springer, New York, USA, 2 edition.

Rao, C. V. (2000). *Moving Horizon Strategies for the Constrained Monitoring and Control of Nonlinear Discrete-Time Systems.* PhD thesis, University of Wisconsin Madison.

Rao, C. V., Rawlings, J. B., and Lee, J. H. (2001). Constrained linear state estimation – a moving horizon approach. *Automatica*, pages 1619–1628.

Schön, T., Gustafsson, F., and Hansson, A. (2003). A note on state estimation as a convex optimization problem. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 6, pages 61–64, Hong Kong.

Schön, T. B. (2008). Nonlinear state estimation – an engineering perspective. Lecture Notes in Dynamic Vision.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.