

# 4

## Modellierung

In diesem Kapitel werden wir uns mit der Modellierung von Bildverarbeitungsaufgaben als Variationsprobleme beschäftigen. Wir gehen dabei von den gegebenen Daten  $f$  (bzw.  $F$  im diskreten Fall) aus und suchen eine Funktion  $u$  (bzw. wieder diskret  $U$ ) als Lösung. Wir werden sehen, dass wir immer Probleme der Form

$$J(u) = D(u, f) + \alpha R(u) \quad (4.1)$$

erhalten, wobei  $D$  ein Datenterm und  $R$  ein Regularisierungsfunktional ist. Im einfachsten Fall ist  $u$  ein Bild (Rekonstruktion, Entrauschung), es kann sich aber auch um eine Indikatorkfunktion (im Fall der Segmentierung), eine Deformation (Registrierung), oder ein Geschwindigkeitsfeld (Dynamik) handeln. In der allgemeinen Modellierung lassen sich diese Fälle noch gemeinsam behandeln, wenn wir aber zu Beispielen von Daten- und Regularisierungstermen übergehen werden wir hier eine Unterscheidung vornehmen.

### 4.1 Regularisierungstheorie und MAP-Schätzung

Im wesentlichen kann man zwei verschiedene Zugänge unterscheiden um (4.1) herzuleiten, einen basierend auf deterministischen und einen basierend auf stochastischen Argumenten. Der deterministische Zugang der Regularisierungstheorie nimmt an, dass man eigentlich exakte Daten  $f_*$  reproduzieren möchte, das eigentlich auf eine Gleichung der Form

$$D(u, f_*) = 0 \quad (4.2)$$

führt, wobei  $D$  ein nichtnegatives Datenfunktional ist, dessen Minimum Null genau dann erreicht wird, wenn die Daten exakt reproduziert werden. Die Gleichung ist

damit auch äquivalent zur Minimierung von  $D(u, f_*)$  bezüglich  $u$ . Wir illustrieren dies an einigen Beispielen:

- Im Fal der Entrauschung wollen wir eigentlich die Gleichung  $u = f_*$  lösen, dies können wir durch Minimierung verschiedener Funktionale, z.B. Der einfachen Normen

$$D(u, f_*) = \frac{1}{p} \int_{\Omega} |u - f_*|^p dx \quad (4.3)$$

erreichen.

- Bei der Bildrekonstruktion haben wir eigentlich  $Ku = f_*$  mit  $K$  einem Operator, der in einem Banachraum abbildet. Es ist dann natürlich die Norm in diesem Banachraum, bzw. eine Potenz davon zu minimieren, d.h.

$$D(u, f_*) = \|Ku - f_*\|^p. \quad (4.4)$$

- Bei der Registrierung wollen wir eine Transformation von  $f_R$  auf  $f_T$  finden (die wir wegen der Einheitlichkeit hier  $u$  statt  $y$  nennen), d.h. eigentlich die Gleichung  $f_R = f_T \circ u$  lösen. Auch dies können wir natürlich durch Minimierung von Normen wie etwa

$$D(u, (f_R, f_T)) = \frac{1}{p} \int_{\Omega} |f_R(x) - f_T(u(x))|^p dx. \quad (4.5)$$

In der praktischen Anwendung hat man nun aber nicht exakte, sondern gestörte Daten  $f$ , sodass die alleinige Minimierung von  $D(u, f)$  nicht sinnvoll ist. Insbesondere hat man bei schlecht gestellten Problemen, etwa Rekonstruktion mit kompaktem Operator  $K$ , oft die Existenz eines Minimums von  $D(u, f)$  im Fall gestörter Daten nicht gesichert, selbst wenn es existiert kann es zu unvernünftigen Lösungen führen. Deshalb benutzt man a-priori Information, die eine vernünftige Lösung auszeichnen und modelliert dies mathematisch durch Regularisierungsfunktionale  $R$ , die nicht zu gross werden sollen (umso kleiner  $R$ , umso plausibler die Lösung). Zusätzlich lässt man zu, dass die gestörten Daten nicht exakt reproduziert werden, sondern nur approximiert, d.h. man akzeptiert Lösungen mit

$$D(u, f) \leq \delta, \quad (4.6)$$

mit kleinem  $\delta$  abhängig von der geschätzten Grösse der Störung. Damit kann man im Fall gestörter Daten ein beschränktes Optimierungsproblem formulieren, nämlich die Minimierung von  $R$  unter allen  $u$ , die (4.6) erfüllen. Nun kann man zeigen, dass ein Lagrange-Parameter  $\lambda > 0$  existiert, sodass die beschränkte Optimierung äquivalent ist zur unbeschränkten Optimierung von

$$R(u) + \lambda(D(u, f) - \delta) \rightarrow \min_u. \quad (4.7)$$

Definiert man  $\alpha = \frac{1}{\lambda}$ , dann ist dieses Problem genau äquivalenz zur Minimierung von (4.1). Die Modellierung der Störung bestimmt meist auch den verwendeten Datenterm, da die Annahme, dass für die exakte Lösung  $u_*$  auch wirklich  $D(u_*, f) \leq \delta$  gilt, gerechtfertigt sein sollte. Im Fall der Entrauschung sollte man etwa die Norm so anpassen, dass die Störung in dieser Norm gemessen klein ist, es gilt ja die Dreiecksungleichung

$$\|u_* - f\| \leq \|u_* - f_*\| + \|f - f_*\| = \|f - f_*\|. \quad (4.8)$$

Für typische Rauschmodelle ist also  $p = 1$  oder  $p = 2$  sinnvoll, während grosses  $p$  (insbesondere  $p = \infty$ ) oder auch Sobolev-Normen nicht auf kleine Störungen führen werden. Im Fall eines additiven Gauss'schen Rauschens haben wir bereits gesehen, dass wir in natürlicher Weise eine Annahme der Form

$$\int_{\Omega} (f - f_*)^2 dx \leq \delta \quad (4.9)$$

machen können, wobei  $\delta$  der Varianz des Rauschens entspricht, also scheint die  $L^2$ -Norm naheliegend.

Ein interessantes stochastisches Konzept sind Bayes<sup>6</sup>-sche Modelle. Dabei wird jedem Bild eine a-priori Wahrscheinlichkeit oder Wert einer Wahrscheinlichkeitsdichte zugeordnet. Zusätzlich benötigt man noch eine Wahrscheinlichkeit für das Auftreten einer bestimmten Messung unter einem gegebenen Bild. Das Bayes'sche Modell basiert dann auf der sogenannte *a-posteriori Wahrscheinlichkeit* (a-posteriori probability, auch Bayes'sche Formel genannt)

$$\mathbb{P}(U | F) = \frac{\mathbb{P}(F | U)\mathbb{P}(U)}{\mathbb{P}(F)}. \quad (4.10)$$

Die Formel für  $\mathbb{P}(U | F)$  beschreibt die Wahrscheinlichkeitsdichte, dass die Beobachtung  $F$  von einem Bild  $U$  erzeugt wurde. Dabei ist  $\mathbb{P}(F | U)$  die Wahrscheinlichkeit, dass bei gegebenem Bild  $U$  die Beobachtung  $F$  eintritt, und  $\mathbb{P}(U)$ ,  $F$  sind a-priori Wahrscheinlichkeiten für das Bild und die Beobachtung. Man modelliert also hier sogar eine sehr hochdimensionale Dichte, die die Güte einer möglichen Lösung  $U$  beschreiben kann.

In der Praxis ist natürlich die hochdimensionale Dichte (in einem Raum dessen Dimension der Anzahl der Pixel im Bild entspricht) nicht berechenbar, und wie immer in der Statistik muss man aus der Wahrscheinlichkeit Schätzer konstruieren, der einfach berechenbar ist. Durch Maximierung der Wahrscheinlichkeit kann man einen

---

<sup>6</sup>Thomas Bayes (1702-1761), englischer Mathematiker und Pfarrer. Wurde vor allem durch eine zwei Jahre nach seinem Tod veröffentlichte Arbeit zur Wahrscheinlichkeit bekannt, cf. [?]

solchen Schätzer konstruieren, den sogenannten *maximum a-posteriori probability (MAP)* Schätzer

$$\hat{U} = \arg \max_U \mathbb{P}(U \mid F). \quad (4.11)$$

Man sieht aus (4.10) und (4.11), dass nur die a-priori Wahrscheinlichkeit für  $U$  wichtig ist, da  $F$  ohnehin fix gehalten wird. Der Nenner  $\mathbb{P}(F)$  ist dann nur eine Normierungskonstante. Wir verdeutlichen dieses Modell anhand eines Gauss'schen Modells für das Rauschen und eines sogenannten Gibbs-Modells für die a-priori Wahrscheinlichkeit. In diesem Fall haben wir kontinuierliche Zufallsvariable und verwenden daher die Dichten

$$\rho_{\mathbb{P}}(U \mid F) = \rho_{\mathbb{P}}(F \mid U)\rho_{\mathbb{P}}(U),$$

und den MAP-Schätzer

$$\hat{U} = \arg \max_U \rho_{\mathbb{P}}(U \mid F). \quad (4.12)$$

Da beim Gauss'schen Modell für das Rauschen  $F$  durch Addition normalverteilter Zufallsvariable zu  $U$  entsteht, ist das Ereignis der Beobachtung von  $F$  unter  $U$  äquivalent zu  $n_{ij}^\delta = F_{ij} - U_{ij}$ . Da wir unabhängige normalverteilte Zufallsvariable haben, kann die Wahrscheinlichkeitsdichte als

$$\rho_{\mathbb{P}}(F \mid U) = \prod_{i,j} \frac{1}{\sigma\sqrt{\pi}} e^{-\frac{(F_{ij}-U_{ij})^2}{2\sigma^2}}$$

berechnet werden. Zusammen mit dem Gibbs-Modell erhalten wir die a-posteriori Wahrscheinlichkeitsdichte als

$$\rho_{\mathbb{P}}(U \mid F) = C e^{-\beta E(u)} \prod_{i,j} e^{-\frac{(F_{ij}-U_{ij})^2}{2\sigma^2}}.$$

Den MAP-Schätzer erhalten wir analog durch Minimierung des negativen Logarithmus der Wahrscheinlichkeitsdichte, d.h.

$$\hat{U} = \arg \min_U \left( \frac{1}{2} \sum_{ij} (F_{ij} - U_{ij})^2 + \sigma^2 \beta E(u) \right). \quad (4.13)$$

Bei geeigneter Skalierung (Multiplikation mit Fläche der Pixel) erhält man einen kontinuierlichen Grenzwert als

$$\frac{1}{N_1 N_2} \sum_{ij} (f(x_{ij}) - u(x_{ij}))^2 \rightarrow \int_{\Omega} (f(x) - u(x))^2 dx,$$

für  $N_1, N_2 \rightarrow \infty$  und damit asymptotisch das Variationsmodell

$$\hat{u} = \arg \min_u \left( \frac{1}{2} \int_{\Omega} (u - f)^2 dx + \alpha R(u) \right), \quad (4.14)$$

wobei  $R$  ein passender Grenzwert des Regularisierungsteils ist.

Weitere interessante Beispiele, die man aus der stochastischen Modellierung des Rauschens erhält, sind

- *Additives Laplace-Rauschen:*  $F_{ij} - U_{ij}$  unabhängig verteilt nach Laplace-Verteilung. Diesen Fall nimmt man, wenn auch grössere Abweichungen häufiger auftreten (durch den schnellen Abfall der Gauss-Verteilung sind diese bei additivem Gauss'schen Rauschen sehr selten). Asymptotisch erhält man daraus

$$D(u, f) = \int_{\Omega} |u - f| \, dx. \quad (4.15)$$

- *Poisson-Rauschen:*  $F_{ij}$  is Poisson-Verteilt mit Mittelwert  $U_{ij}$ . Hier erhält man asymptotisch

$$D(u, f) = \int_{\Omega} (u - f \log u) \, dx, \quad (4.16)$$

das man durch Hinzufügen von Konstanten auch als die (nichtnegative) *Kullback-Leibler Divergenz*

$$D(u, f) = \int_{\Omega} (f \log \frac{f}{u} - f + u) \, dx, \quad (4.17)$$

schreiben kann. Das Poisson-Modell entsteht üblicherweise durch einen Zählprozess, etwa wenn man die endliche Anzahl von detektierten Photonen modellieren will.

- *Multiplikatives Rauschen:*  $F_{ij} = n_{ij}^{\delta} U_{ij}$  mit unabhängig verteilten  $n_{ij}^{\delta}$  (man findet dafür verschiedene Verteilungsannahmen). Da das Rauschen nun dem Quotienten  $\frac{F_{ij}}{U_{ij}}$  entspricht, führt dies auf Funktionale, die von  $\frac{f}{u}$  abhängen.

## 4.2 Regularisierungsfunktionale

In diesem Abschnitt werden wir uns zunächst näher mit der Frage beschäftigen, welche Art von Regularisierungsterm ( $F$ ) bzw. welchen Funktionenraum wir für das Entrauschen von Bildern sinnvollerweise benutzen können. Die Lebesgue-Räume  $L^p(\Omega)$  sollten alle sinnvollen Bilder enthalten, da aber auch Rauschen in diesen Räumen liegen kann (insbesondere das Gauss'sche Rauschen in unserem Modell in  $L^2(\Omega)$ ) sind sie aber zu gross gewählt - man kann in  $L^p(\Omega)$  nicht zwischen Signal und Rauschen unterscheiden. Besonders naheliegend ist deshalb die Betrachtung der Sobolev-Räume  $W^{1,p}(\Omega)$ ,  $p \geq 1$ . Diese sind Teilräume der Lebesgue-Räume  $L^p(\Omega)$ ,

die auch noch einen  $p$ -integrierbaren Gradienten haben sollen und damit Oszillationen (Rauschen) zu einem signifikant höheren Wert der Norm führen. Andererseits könnte die Wahl eines solchen Sobolev-Raums zu einschränkend sein, um interessante Bilder (mit Kanten) zuzulassen. Dies ist für  $p > 1$  auch der Fall, wie wir aus den folgenden Resultaten sehen.

**Lemma 4.2.1.** *Sei  $u \in W^{1,p}(\Omega)$ ,  $p > 1$  wobei  $\Omega \subset \mathbb{R}^1$  ein Intervall sei. Dann ist  $u$  stetig.*

**Beweis.** Sei zunächst  $u \in C^1(\Omega) \subset W^{1,p}(\Omega)$ . Dann folgt

$$|u(y) - u(x)| = \left| \int_x^y u'(z) dz \right| \leq \left( \int_x^y |u'(z)|^p dz \right)^{1/p} |x - y|^{1/p'},$$

wobei  $1/p + 1/p' = 1$  gilt. Also folgt

$$|u(y) - u(x)| \leq \|u\|_{W^{1,p}(\Omega)} |x - y|^{1/p'}.$$

Für beliebiges  $u \in W^{1,p}(\Omega)$  können wir nun die Dichtheit von  $C^1(\Omega)$  in  $W^{1,p}(\Omega)$  verwenden, d.h. es existiert eine Folge  $u_n \in C^1(\Omega)$  die gegen  $u$  in der Norm von  $W^{1,p}(\Omega)$  und auch punktweise konvergiert. Damit folgt fast überall

$$|u(y) - u(x)| = \lim |u_n(x) - u_n(y)| \leq \lim \|u_n\|_{W^{1,p}(\Omega)} |x - y|^{1/p'} = \|u\|_{W^{1,p}(\Omega)} |x - y|^{1/p'}.$$

Damit ist  $u$  sogar Hölder-stetig mit Exponent  $1/p'$ .  $\square$

**Lemma 4.2.2.** *Sei  $D \subset \Omega$  ein Gebiet mit  $C^1$ -Rand. Dann ist die Funktion*

$$u(x) = \begin{cases} 1 & \text{für } x \in D \\ 0 & \text{sonst} \end{cases}$$

nicht in  $W^{1,p}(\Omega)$  für  $p \geq 1$ .

**Beweis.** Der distributionelle Gradient einer Funktion  $u$  ist definiert durch das lineare Funktional

$$\langle \nabla u, \varphi \rangle = - \int_{\Omega} u(\nabla \cdot \varphi) dx \quad \forall \varphi \in C_0^\infty(\Omega; \mathbb{R}^d).$$

Für eine Funktion der obigen Gestalt folgt dann mit dem Gauss'schen Satz

$$\langle \nabla u, \varphi \rangle = - \int_D \nabla \cdot \varphi dx = - \int_{\partial D} \varphi \cdot n d\sigma.$$

Da dieses lineare Funktional auf  $\partial D$ , d.h. einer Nullmenge des Lebesgue-Maßes konzentriert ist, kann es für kein  $p' \in [1, \infty]$  zu einem linearen Funktional in  $L^{p'}(\Omega, \mathbb{R}^d)$

erweitert werden. Damit liegt der distributionelle Gradient auch nicht in  $L^p(\Omega, \mathbb{R}^d) \subset L^{p'}(\Omega, \mathbb{R}^d)^*$ .  $\square$

Als Folgerung sehen wir also, dass  $W^{1,p}$  für  $p > 1$  sicher nicht in Frage kommt, da Unstetigkeiten (Kanten) nicht zugelassen werden. Auch der Fall  $p = 1$  würde nach dem obigen Argument stückweise konstante Funktionen nicht zulassen. Darüber hinaus ergibt sich in  $W^{1,1}(\Omega)$  eine funktionalanalytische Schwierigkeit, da dieser Raum (analog zu  $L^1(\Omega)$ ) nicht der Dualraum eines Banachraumes ist. Man kann also keine schwach-\* Topologie verwenden und damit auch den Satz von Banach-Alaoglu nicht anwenden. In beiden Fällen liegt das Problem darin, dass gewisse Maße wie zum Beispiel Dirac  $\delta$ -Distributionen nicht als Gradient einer Funktion  $W^{1,1}(\Omega)$  zugelassen werden, aber durch  $W^{1,1}$ -Funktionen mit beschränkter Norm approximiert werden können.

Um einen sinnvolleren Funktionenraum für Bilder zu erhalten, muss  $W^{1,1}(\Omega)$  noch einmal vergrößert werden. Dabei geht man analog vor wie beim "Übergang von  $L^1(\Omega)$  zum Raum der Radon-Maße. Wir definieren deshalb den Raum  $BV(\Omega)$  von Funktionen beschränkter (totaler) Variation

$$BV(\Omega) = \{u \in L^1(\Omega) \mid |u|_{BV} < \infty\}, \quad (4.18)$$

wobei die totale Variation gegeben ist durch

$$|u|_{BV} := \sup_{\varphi \in C_0^\infty(\Omega; \mathbb{R}^d), \|\varphi\|_\infty \leq 1} \int_{\Omega} u \nabla \cdot \varphi \, dx. \quad (4.19)$$

Die Norm in  $BV(\Omega)$  ist gegeben durch

$$\|u\|_{BV} := \|u\|_{L^1} + |u|_{BV}. \quad (4.20)$$

Die Wahl des Raums  $C_0^\infty(\Omega; \mathbb{R}^d)$  in der Definition ist nicht essentiell. Wegen der Dichte dieses Raums in anderen Räumen erhalten wir das selbe Supremum in jedem  $C_0^k(\Omega; \mathbb{R}^d)$  für  $k \geq 0$ . Wir werden im folgenden auch diese unterschiedlichen Räume in der Definition je nach Notwendigkeit einsetzen. Für Funktionen in  $u \in W^{1,1}(\Omega)$  rechnet man leicht nach, dass  $\|u\|_{BV} = \|u\|_{W^{1,1}}$  gilt. Damit enthält  $BV(\Omega)$  auf jeden Fall  $W^{1,1}$ . Man sieht auch leicht, dass  $BV(\Omega)$  auch unstetige Funktionen enthält. Für die stückweise konstante Funktion  $u$  aus Lemma 4.2.2 gilt

$$\begin{aligned} |u|_{BV} &= \sup_{\varphi \in C_0^\infty(\Omega; \mathbb{R}^d), \|\varphi\|_\infty \leq 1} \int_D \nabla \cdot \varphi \, dx \\ &= \sup_{\varphi \in C_0^\infty(\Omega; \mathbb{R}^d), \|\varphi\|_\infty \leq 1} \int_{\partial D} \varphi \cdot n \, d\sigma \\ &= \int_{\partial D} d\sigma < \infty. \end{aligned}$$

Sobald also  $\partial D$  ein endliches  $d - 1$ -dimensionales Hausdorff-Maß (bei Kurven also einfach endliche Länge) hat, ist die totale Variation der stückweise konstanten Funktion endlich. Wir sehen auch dass die totale Variation in diesem Fall gleich der Länge der Kurve ist, d.h. eine Minimierung (bzw. zumindest Verkleinerung) der totalen Variation sollte auch die Kurven glätten.

Nach dieser Argumentation ist es naheliegend geglättete Funktionen im Raum beschränkter Variation zu suchen, bzw. die totale Variation sogar als Regularisierungsfunktional zu verwenden. Damit erhält man das Rudin-Osher-Fatemi (ROF) Modell zum Entrauschen, das in der Minimierung

$$J(u) = \frac{\lambda}{2} \int_{\Omega} (u - f)^2 dx + |u|_{BV} \rightarrow \min_{u \in BV(\Omega)} \quad (4.21)$$

besteht. Dieses Modell und Variationen davon sind heute eines der populärsten Forschungsgebiete in der Bildverarbeitung. Wie wir sehen werden, erhält man daraus (im Gegensatz zu vielen anderen Variationsmethoden und Filtern) scharfe Kanten bei der Rekonstruktion und auch recht klare geometrische Interpretationen bezüglich der Isokonturen. Auch für die Segmentierung können wir die totale Variation als Regularisierung verwenden, da ja die totale Variation einer Indikatorfunktion gleich dem Perimeter der Kantenmenge ist.

In manchen Fällen bringt man a-priori Information dadurch ein, indem man ein Dictionary möglicher Bildstrukturen zur Verfügung hat, also ein System von Funktionen  $\{\varphi_j\}_{j=0}^{\infty}$  und das Bild sich als Linearkombination möglichst weniger  $\varphi_j$  schreiben lassen soll (Sparsity). Auch hier sieht man leicht, dass man mit strikt konvexen und differenzierbaren Funktionalen dieses Ziel nicht erreicht. So erhält man etwa bei quadratischen Regularisierungen viele kleine Koeffizienten in den Linearkombinationen, aber sehr wenige gleich Null. Ideal wäre eine  $\ell^0$ -Norm als Regularisierung, also eine Bestrafung mit der Anzahl der Nichtnullelemente. Dieses Funktional ist jedoch extrem nichtkonvex, sodass eine Berechnung globaler Minima schwierig ist. Deshalb verwendet man wieder eine konvexe Regularisierung mit einer  $\ell^1$ -Norm, d.h.

$$J(u) = \sum_j |c_j|, \quad \text{wobei } u = \sum_j c_j \varphi_j. \quad (4.22)$$

Eine zweite häufig verwendete Form (die zur obigen äquivalent ist, wenn  $\{\varphi_j\}$  eine Orthonormalbasis ist) ist die  $\ell^1$ -Norm der Skalarprodukte

$$J(u) = \sum_j |\langle u, \varphi_j \rangle|. \quad (4.23)$$

Die obigen Argumente gelten nicht für die Deformationen bei der Regularisierung oder die Vektorfelder in dynamischen Problemen. Hier will man sehr oft glatte

---

Lösungen, weshalb die Verwendung einfacherer quadratischer Regularisierungsfunktionale durchaus angebracht ist. Hier ist etwa das Quadrat der  $W^{1,2}$ -Norm ein gängiges Regularisierungsfunktional.