Teil IX

Modellbasierte Optimierung

9 Zufallsfeld

Dieses Kapitel ist als Nachtrag zu Kapitel 3 gedacht.

- Wir werden modellbasierte Optimierung kennenlernen.
- Idee ist es, teure Auswertungen einer unbekannten Zielfunktion (black box) möglichst zu minimieren.
- Zunächst jedoch einige Grundlagen.

9.1 Zufallsfeld

- Ein *stochastischer Prozess* ist eine Familie von Zufallsvariablen $\{Z_i \mid i \in I\}$ mit $Z_i : (\Omega, \mathcal{A}, P) \mapsto (E, \mathcal{G})$, wobei $I \neq \emptyset$ eine Indexmenge, $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und $(E, \mathcal{G})$ ein Zustandsraum ist [115].
- Die Indexmenge wird z.B. bei zeitlichen Prozessen als $I \subseteq \mathbb{Z}$ oder $I \subseteq \mathbb{N}$ im diskreten oder als $I \subset \mathbb{R}$ im kontinuierlichen Fall festgelegt.
- Bei räumlichen Prozessen ist die Indexmenge häufig eine Teilmenge des $\mathbb{R}^d$ ($d \geq 1$) und der entsprechende Prozess wird auch als Zufallsfeld bezeichnet.
- Im Folgenden betrachten wir Zufallsfelder (ZF) der Form

$$\left\{ Z(x) \mid x \in \mathbb{D} \subseteq \mathbb{R}^d, d \geq 1 \right\}. \tag{1}$$

- Ein ZF ist ein *Gauß'sches ZF*, wenn für beliebige $x_1, \ldots, x_n \in \mathbb{D}$ der Vektor der Zufallsvariablen $Z = (Z(x_1), \ldots, Z(x_n))^T$ multivariat normalverteilt ist und sich damit vollständig durch $E[Z(x)]$ und $Cov[Z(x), Z(x')]$ charakterisieren lässt [109].

9.1 Zufallsfeld

- Ein ZF ist schwach stationär (mittelwert- und kovarianzstationär) wenn die beiden ersten Momente existieren und es gilt $\forall x, x' \in \mathbb{D}$:

$$E\left[Z\left(x\right)\right] = \mu \tag{2}$$

$$\begin{aligned} \kappa\left(x, x'\right) &= Cov\left[Z\left(x\right), Z\left(x'\right)\right] = Var\left[Z\left(x\right)\right] Cor\left[Z\left(x\right), Z\left(x'\right)\right] \\ &= \sigma^2 \rho\left(x - x', \psi\right) \end{aligned} \tag{3}$$

  Damit hängt die räumliche Variabilität des ZF nur von der Differenz $x - x'$ und der Form der Korrelationsfunktion ρ, sowie deren Parametern $\psi = \left(\psi_1, \ldots, \psi_d\right)^T$ mit $\psi_j > 0 \;\; \left(j = 1, \ldots, d\right)$ ab [140].
- Annahme: Eigenschaften des Prozesses über alle $x \in \mathbb{D}$ hinweg konstant.
- Bei Gaußprozess folgt aus schwacher auch die starke Stationarität, oder hier einfach Stationarität.
- Ist die Kovarianzfunktion

$$\kappa\left(\delta\right) = \kappa\left(x, x'\right) \quad \forall x, x' \in \mathbb{D} : \left\|x - x'\right\| = \delta \tag{4}$$

  nur vom Euklidischen Abstand $\left\|\cdot\right\|$ zweier Punkte abhängig, so ist es ein isotropes ZF und damit rotationsinvariant.
- Damit sind stationäre isotrope ZFs sind bewegungsinvariant [173, 64].

9.1 Zufallsfeld: Kovarianz- und Korrelationsfunktion

Unter Berücksichtigung der schwachen Stationarität und Isotropie gilt für die Kovarianzfunktion κ:

($\kappa 1$) $\kappa(x,x) = \sigma^2 \rho(x-x,\psi) = \sigma^2 = Var[Z(x)] \geq 0 \ \forall x \in \mathbb{D}$

($\kappa 2$) $\kappa(x,x') = \kappa(x',x) \ \forall x,x' \in \mathbb{D}$ (Symmetrie)

($\kappa 3$) $|\kappa(x,x')| \leq \kappa(x,x) \ \forall x,x' \in \mathbb{D}$ (Beschränktheit)

($\kappa 4$) κ ist positiv semidefinit.

Da $\rho(x-x',\psi) = \frac{\kappa(x,x')}{\kappa(x,x)}$ mit $\rho(x-x,\psi) = \rho(0,\psi) = 1$, sofern $\kappa(x,x) > 0$, gelten die Eigenschaften ($\kappa 2$), ($\kappa 3$) und ($\kappa 4$) entsprechend auch für die Korrelationsfunktion.

9.1 Zufallsfeld: Kovarianz- und Korrelationsfunktion

Mögliche Wahl der Korrelationsfunktion:

- Exponentielle Korrelationsfunktion:

$$\rho\left(x - x', \psi\right) = \exp\left(-\left\|x - x'\right\|_{\Delta}\right) \tag{5}$$

  - Distanz zwischen x und x' wird über den Abstand $\left\|x - x'\right\|_{\Delta} = \sqrt{\left(x - x'\right)^T \Delta^{-1} \left(x - x'\right)}$ mit Diagonalmatrix $\Delta = \text{diag}\left(\psi_1^2, \ldots, \psi_d^2\right)$ und $\psi_1 = \ldots = \psi_d$ gemessen.
  - Da Parameter über alle Dimensionen gleich sind, hängt die Korrelationsfunktion und somit die Kovarianzfunktion nur vom Abstand der Punkte ab und ist damit isotrop.
- anisotrope Matérn 5/2 Korrelationsfunktion:

$$\rho\left(x - x', \psi\right) = \left[1 + \sqrt{5}\left\|x - x'\right\|_{\Delta} + \frac{5}{3}\left\|x - x'\right\|_{\Delta}^2\right] \exp\left(-\sqrt{5}\left\|x - x'\right\|_{\Delta}\right) \tag{6}$$

mit $\left\|x - x'\right\|_{\Delta} = \sqrt{\left(x - x'\right)^T \Delta^{-1} \left(x - x'\right)}, \Delta = \text{diag}\left(\psi_1^2, \ldots, \psi_d^2\right)$.

  - Der Skalierungsparameter ψ legt fest, wie schnell die Korrelation bzw. Kovarianz zwischen zwei Punkten mit steigender Distanz gegen Null läuft.
  - Daher wird ψ oftmals als Reichweite interpretiert.

9.1 Zufallsfeld: Kovarianz- und Korrelationsfunktion

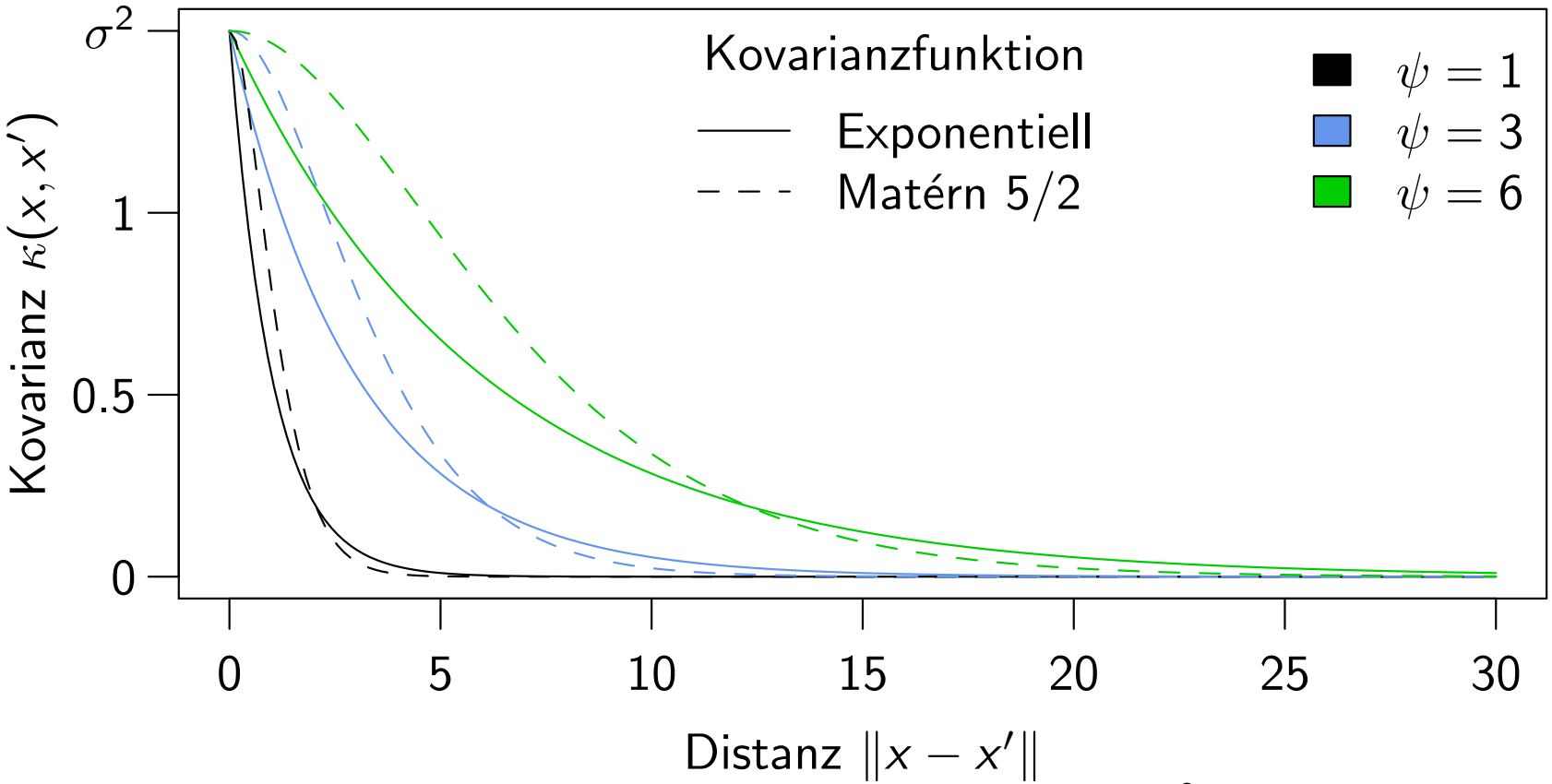


Abb. 9.1 : Matérn 5/2 und exponentielle Kovarianzfunktion mit $\sigma^2 = 1.5$ und unterschiedlichen Reichweiten ψ für $d = 1$.

9.1 Zufallsfeld für verrauschte Beobachtungen

Zufallsfeld mit unkorreliertem Anteil: $Z_\varepsilon(x) = Z(x) + \varepsilon(x)$

- $E[\varepsilon(x)] = 0$
- $Cov[\varepsilon(x), \varepsilon(x')] = \sigma_\varepsilon^2 \mathbb{1}_{x=x'}(x, x')$
- $Cov[Z(x), \varepsilon(x')] = 0$
- $Cov[Z_\varepsilon(x), Z_\varepsilon(x')] = \kappa_\varepsilon(x, x') = \sigma^2 \rho(x - x', \psi) + \sigma_\varepsilon^2 \mathbb{1}_{x=x'}(x, x')$

Die Varianz von $Z_\varepsilon(x)$ ist somit um σ_ε^2 größer als die von $Z(x)$.

Aus der nächsten Abbildung ist zu sehen, dass sich für $Z(x)$ bzw. $Z_\varepsilon(x)$ je nach Wahl der Parameter viele unterschiedliche Prozessverläufe ereignen können.

9.1 Zufallsfeld für verrauschte Beobachtungen

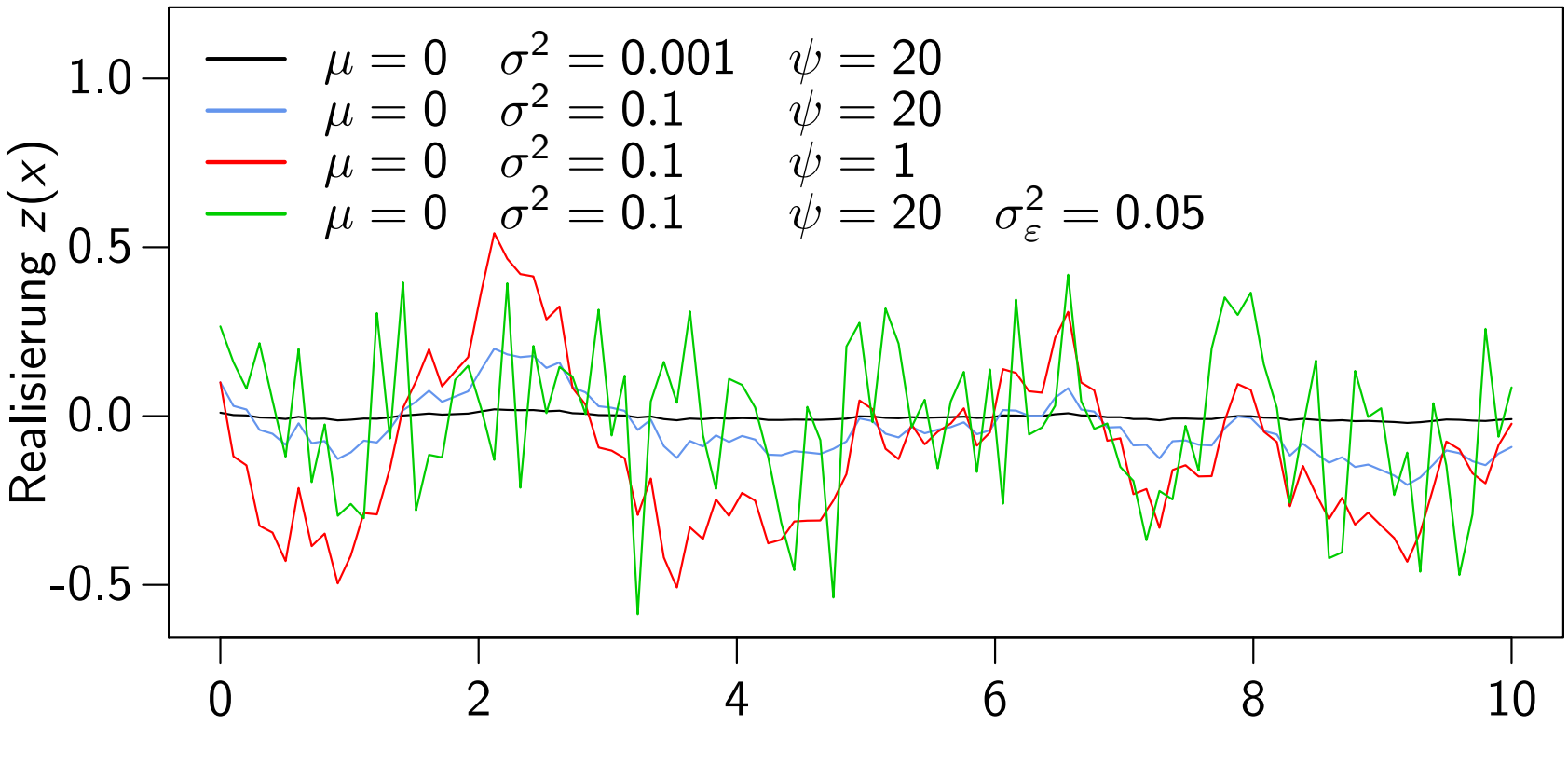


Abb. 9.2 : Realisierungen des Zufallsfeldes $\{Z(x) \mid x \in [0, 10]\}$ für unterschiedliche Modellparameter mit exponentieller Kovarianzfunktion.

9.1 Zufallsfeld: Schätzung der Parameter

Um aus den Realisierungen $z_\varepsilon(x_1), \ldots, z_\varepsilon(x_n)$ die Parameter zu schätzen, wird

- das *Variogramm*

$$\begin{aligned} \nu(\delta) = \nu(x, x') &= Var[Z_\varepsilon(x) - Z_\varepsilon(x')] \\ &= Var[Z_\varepsilon(x)] + Var[Z_\varepsilon(x')] - 2Cov[Z_\varepsilon(x), Z_\varepsilon(x')] \\ &= 2\sigma^2[1 - \rho(x - x', \psi)] + 2\sigma_\varepsilon^2[1 - \mathbb{1}_{x=x'}(x, x')] \end{aligned} \tag{7}$$

- bzw. das *Semivariogramm* $\nu_{\text{semi}}(\delta) = \frac{1}{2}\nu(\delta)$ herangezogen.
- Dabei entspricht $\lim_{\delta \to \infty} \nu_{\text{semi}}(\delta)$ der Prozessvarianz $\sigma^2 + \sigma_\varepsilon^2$ und $\lim_{\delta \to 0} \nu_{\text{semi}}(\delta)$ der Fehlervarianz σ_ε^2 (*Nuggetvarianz* in der Geostatistik).

9.1 Zufallsfeld: Schätzung der Parameter

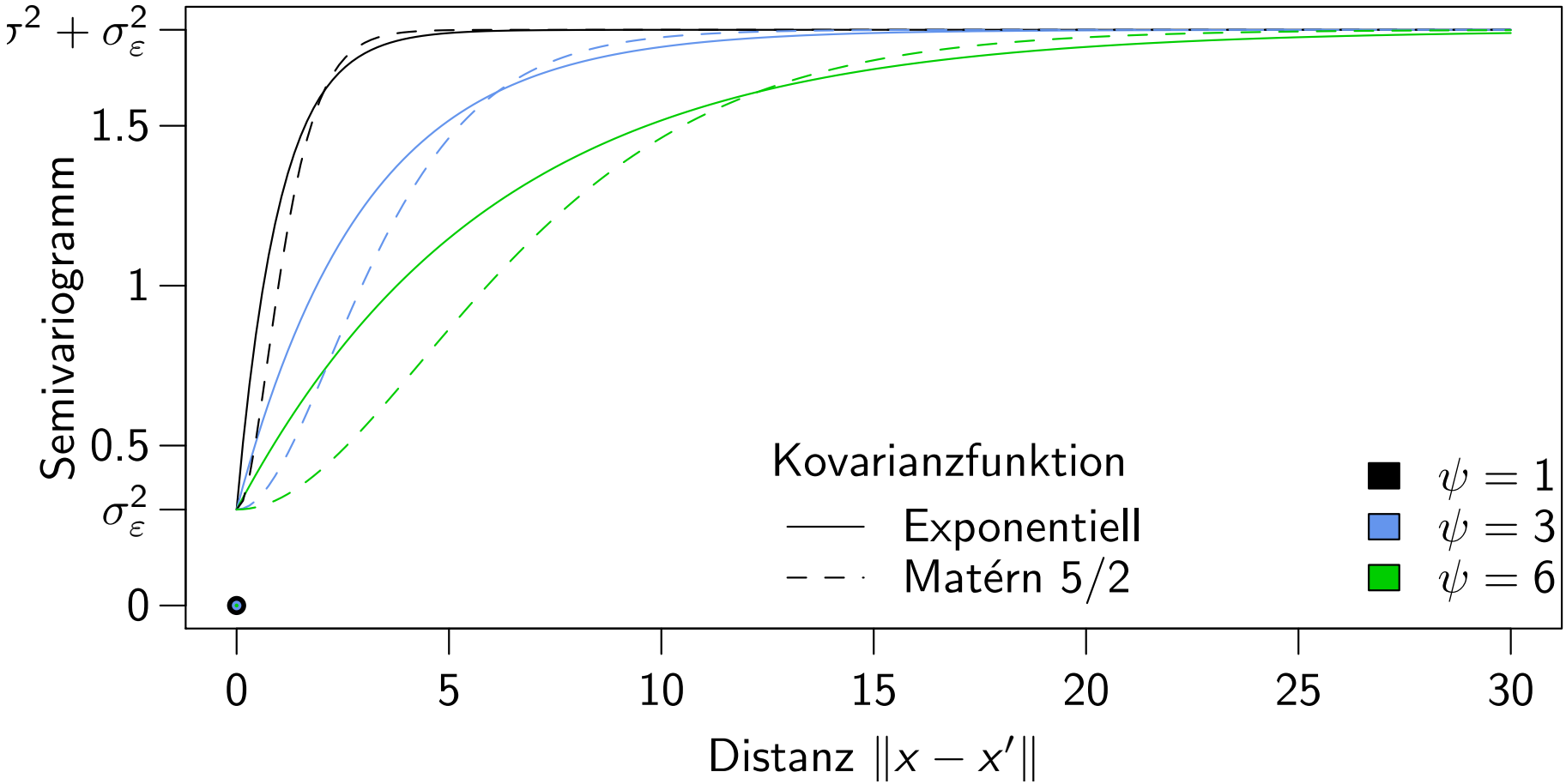


Abb. 9.3 : Theoretisches Semivariogramm $\nu_{\text{semi}}(x, x')$ mit $\sigma^2 = 1.5$ und $\sigma_\varepsilon^2 = 0.3$.

9.1 Zufallsfeld: Schätzung der Parameter

- Der Momentenschätzer für das Semivariogramm einer Punktdistanz δ ist gegeben durch

$$\widehat{\nu}_{\text{semi}}(\delta) = \frac{1}{2|N(\delta)|} \sum_{(x,x')\in N(\delta)} (z_\varepsilon(x) - z_\varepsilon(x'))^2, \tag{8}$$

  wobei

$$N(\delta) = \{(x, x') \mid \|x - x'\| = \delta\} \tag{9}$$

  die Menge der Punkte mit Abstand δ ist und $|N(\delta)|$ die Anzahl ihrer Elemente.
- Entsprechend ist $\widehat{\nu}(\delta) = 2\widehat{\nu}_{\text{semi}}(\delta)$ der Momentenschätzer für das Variogramm.
- Damit die Anzahl der Elemente in $N(\delta)$ nicht zu klein ausfällt, wird in der Praxis das Semivariogramm für disjunkte Distanzintervalle $\delta = (\delta_l, \delta_u]$ betrachtet:

$$N(\delta) = N((\delta_l, \delta_u]) = \{(x, x') \mid \delta_l < \|x - x'\| \leq \delta_u\} \tag{10}$$

9.1 Zufallsfeld: Schätzung der Parameter

- Die gewichteten Kleinste-Quadrate-Schätzer für σ^2, σ_ε^2 und ψ werden durch das Minimieren der Verlustfunktion

$$L\left(\sigma^2, \sigma_\varepsilon^2, \psi\right) = \sum_{\delta} |N(\delta)| \left(\widehat{\nu}(\delta) - \nu(\delta)\right)^2 \tag{11}$$

ermittelt, wobei die Art der Kovarianzfunktion (z.B. Matérn 5/2) vorgegeben wird.

9.2 Kriging

- Das Kriging-Modell wurde von D. G. Krige, einem Bergbauingenieur aus Südafrika, 1951 vorgestellt und prägte die Forschung im Bereich der Geostatistik nachhaltig.
- Es liefert den besten linearen unverzerrten Schätzer für einen regionalisierten Gaußprozess gegeben eine Reihe von Beobachtungen.
- Sei $Z(x)$ ein Gaußprozess mit $E[Z(x)] = \mu$ und $Cov[Z(x), Z(x')] = \kappa(x, x')$.
  Dann folgen die Zufallsvariablen $Z = (Z(x_1) \ \cdots \ Z(x_n))^T$ zu den Realisierungen $z = (z(x_1) \ \cdots \ z(x_n))^T$ an den Stellen $x_1, \ldots, x_n \in \mathbb{D} \subseteq \mathbb{R}^d$ der gemeinsamen Normalverteilung $\mathcal{N}(\mu 1_n, K)$ mit Kovarianzmatrix

$$K = \begin{pmatrix} k(x_1)^T \\ \vdots \\ k(x_n)^T \end{pmatrix} \in \mathbb{R}^{n \times n} \tag{12}$$

  bestehend aus den Kovarianzvektoren

$$k(x_i) = [\kappa(x_i, x_1) \ \cdots \ \kappa(x_i, x_n)]^T \in \mathbb{R}^n \quad (i = 1, \ldots, n). \tag{13}$$

9.2 Kriging

- Da $Z(x)$ und Z normalverteilt sind, folgt aus

$$\begin{pmatrix} Z(x) \\ Z \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ 1\mu \end{pmatrix}, \begin{pmatrix} \kappa(x,x) & k(x)^T \\ k(x) & K \end{pmatrix} \right) \tag{14}$$

  die Verteilung des bedingten Prozesses

$$Z(x)\,|Z \sim \mathcal{N}\left(\mu + k(x)^T K^{-1}(Z - 1\mu), \kappa(x,x) - k(x)^T K^{-1} k(x) \right), \tag{15}$$

  wobei wir davon ausgehen, dass μ bekannt ist.
- Da dies in den meisten Fällen nicht zutrifft, muss μ zunächst geschätzt werden, was typischerweise zu einer höheren Prognosevarianz führt.
- Es wird ab hier darauf verzichtet die Abhängigkeit der Kovarianzmatrix bzw. deren Vektoren von den Parametern θ_κ hervorzuheben.

9.2 Kriging: Die Schätzung

- Die unbekannten Parameter der Verteilung μ und θ_κ werden mit der Maximum-Likelihood-Methode geschätzt, wobei die Likelihood

$$\begin{aligned} L(\mu, \theta_\kappa, Z) &= f(Z \mid \mu, \theta_\kappa) \\ &= (2\pi)^{-\frac{n}{2}} (\det K)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(Z - 1_n\mu)^T K^{-1}(Z - 1_n\mu)\right) \end{aligned} \tag{16}$$

maximiert wird. Dabei lässt sich ein geschlossener Ausdruck

$$\widehat{\mu} = \underset{\mu \in \mathbb{R}}{\arg\max}\, L(\mu, \theta_\kappa, Z) = \frac{1_n^T K^{-1} Z}{1_n^T K^{-1} 1_n} \tag{17}$$

für den Schätzer des Erwartungswerts μ in Abhängigkeit von den übrigen Parametern herleiten.

9.2 Kriging: Die Schätzung

- Im nicht verrauschten Prozess mit $K = \sigma^2 R$, wobei R entsprechend zu K die Korrelationsmatrix von Z ist, lässt sich für die Prozessvarianz folgender Ausdruck ermitteln:

$$\widehat{\sigma}^2 = \underset{\sigma^2 \in \mathbb{R}_+}{\arg\max}\, L\left(\widehat{\mu}, \sigma^2, \psi, Z\right) = \frac{1}{n}\left(Z - 1_n \widehat{\mu}\right)^T R\left(\psi\right)^{-1}\left(Z - 1_n \widehat{\mu}\right) \tag{18}$$

  Damit die lineare Vorhersage $\widehat{\lambda}\left(x\right)^T Z$ für $Z\left(x\right)$ an der Stelle x mit Gewichten $\widehat{\lambda}\left(x\right) \in \mathbb{R}^n$ erwartungstreu ist, muss

$$E\left[Z\left(x\right)\right] = \mu \overset{!}{=} \widehat{\lambda}\left(x\right)^T 1_n \mu = E\left[\widehat{\lambda}\left(x\right)^T Z\right] \tag{19}$$

  gelten. Dies ist dann der Fall, wenn die Bedingung $\widehat{\lambda}\left(x\right)^T 1_n = 1$ erfüllt ist.

9.2 Kriging: Die Schätzung

- Um die Gewichte zu schätzen, wird der mittlere quadratische Fehler der Vorhersage

$$MSPE\left(\lambda(x)^T Z\right) = E\left[\left(Z(x) - \lambda(x)^T Z\right)^2\right]$$
$$= \kappa(x,x) - 2\lambda(x)^T k(x) + \lambda(x)^T K\lambda(x) \tag{20}$$

unter der Nebenbedingung $\lambda(x)^T 1_n = 1$ minimiert. Sei dabei c der Lagrange-Multiplikator und

$$L(\lambda(x), c) = \kappa(x,x) - 2\lambda(x)^T k(x) + \lambda(x)^T K\lambda(x) - c\left(\lambda(x)^T 1_n - 1\right) \tag{21}$$

die zu minimierende Funktion mit den Ableitungen

$$\frac{\partial L(\lambda(x), c)}{\partial \lambda(x)} = -2k(x) + 2K\lambda(x) \tag{22}$$

und

$$\frac{\partial L(\lambda(x), c)}{\partial c} = -1_n^T \lambda(x) + 1. \tag{23}$$

9.2 Kriging: Die Schätzung

- Das Nullsetzen der Ableitungen führt zu

$$\widehat{\lambda}(x) = K^{-1}\left(k(x) + \frac{c}{2}1_n\right) \tag{24}$$

und

$$\widehat{c} = 2\frac{1 - 1_n^T K^{-1} k(x)}{1_n^T K^{-1} 1_n} \tag{25}$$

und das Einsetzen von (25) in (24) zur Lösung

$$\widehat{\lambda}(x) = K^{-1}\left(k(x) + \frac{\widehat{c}}{2}1_n\right) = K^{-1}\left(k(x) + 1_n\frac{1 - 1_n^T K^{-1} k(x)}{1_n^T K^{-1} 1_n}\right). \tag{26}$$

9.2 Kriging: Die Schätzung

- Da die Hessematrix $H(\widehat{c}) = \frac{\partial^2 L(\lambda(x),c)}{\partial \lambda(x) \partial \lambda(x)^T}(\widehat{c}) = 2K$ positiv definit ist, aufgrund der Eigenschaften der Kovarianzmatrix K, folgt, dass in

$$\widehat{\lambda}(x) = \underset{\left\{\lambda(x) \in \mathbb{R}^n \mid \lambda(x)^T 1_n = 1\right\}}{\arg\min} MSPE\left(\lambda(x)^T Z\right) \tag{27}$$

  tatsächlich das gesuchte Minimum ist.
- Daraus resultiert die Kriging-Vorhersage

$$\widehat{Z}(x) = \widehat{\lambda}(x)^T Z = \widehat{\mu} + k(x)^T K^{-1} \left[Z - 1_n \widehat{\mu}\right], \tag{28}$$

  welche erwartungstreu ist, denn:

9.2 Kriging: Die Schätzung

- $$\begin{aligned} E\left[\widehat{Z}(x)\right] &= E\left(\widehat{\lambda}(x)^T Z\right) \\ &= E(\widehat{\mu}) + k(x)^T K^{-1}\left[E(Z) - 1_n E(\widehat{\mu})\right] \\ &= \mu + k(x)^T K^{-1}\left[1_n\mu - 1_n\mu\right] = \mu = E\left[Z(x)\right] \end{aligned} \tag{29}$$

  aufgrund der Erwartungstreue des ML-Schätzers für μ

$$E(\widehat{\mu}) = E\left(\frac{1_n^T K^{-1} Z}{1_n^T K^{-1} 1_n}\right) = \frac{1_n^T K^{-1} E(Z)}{1_n^T K^{-1} 1_n} = \frac{1_n^T K^{-1} 1_n \mu}{1_n^T K^{-1} 1_n} = \mu \tag{30}$$

  mit der Varianz

$$Var(\widehat{\mu}) = \frac{1}{1_n^T K^{-1} 1_n}. \tag{31}$$

9.2 Kriging: Die Schätzung

- Die Kriging-Varianz ist gegeben als

$$\begin{aligned}\widehat{s}^{\,2}(x) &= \mathit{Var}\left(\widehat{\lambda}(x)^T Z - Z(x)\right) \overset{(29)}{=} \mathit{MSPE}\left(\widehat{\lambda}(x)^T Z\right) \\ &\overset{(20)}{=} \kappa(x,x) - 2\widehat{\lambda}(x)^T k(x) + \widehat{\lambda}(x)^T K \widehat{\lambda}(x) \\ &= \kappa(x,x) - k(x)^T K^{-1} k(x) + \frac{\left(1 - 1_n^T K^{-1} k(x)\right)^2}{1_n^T K^{-1} 1_n}, \qquad (32)\end{aligned}$$

vgl. [57].

9.2 Kriging: Die Schätzung

Im Fall eines *unverrauschten* Prozesses $Z(x)$ gegeben die *unverrauschten* Beobachtungen aus Z ist die Vorhersage gegeben als $\widehat{Z}(x_i) = z(x_i)$ mit $\widehat{s}^{\,2}(x_i) = 0$ für jedes bereits bekannte $x_i \in \{x_1, \ldots, x_n\}$, da hier $k(x_i)^T$ die $i-$te Zeile der Kovarianzmatrix K ist und damit $k(x_i)^T K^{-1} = e_i^T$ mit e_i $i-$ter Einheitsvektor.

- Dies gilt auch für den Fall, dass $Z_\varepsilon(x)\,|Z_\varepsilon$ modelliert wird, sofern es nicht mehrere Realisationen $z^{(1)}(x), \ldots, z^{(l)}(x)$ für dasselbe x gibt.
  Dieses Modell wird auch als Kriging mit Nugget-Effekt bezeichnet.
- Im Zusammenhang mit der modellbasierten Optimierung ist die Vorhersage eines vielversprechenden Punktes anhand des Modells $Z(x)\,|Z$ zum Standard bei Computerexperimenten geworden.
  Es wird angenommen, dass diese Computerexperimente deterministisch sind.

9.2 Kriging: Die Schätzung

Liegt dem Computerexperiment per Definition eine *stochastische* Komponente zugrunde, so erscheint es unplausibel ein Modell mit interpolierenden Eigenschaften zu verwenden.

- Falls die Beobachtungen Realisationen von Z_ε sind, wir jedoch an dem wahren Prozess $Z(x)$ interessiert sind, verliert die zugehörige Krigingvorhersage ihre interpolierende Eigenschaft.
- In diesem Fall betrachten wir als Vorhersage den Erwartungswert von $Z(x)|Z_\varepsilon$ (noisy Kriging-Modell).
- Hierbei enthält die Kovarianzmatrix K_ε als Einträge die Werte der Kovarianzfunktion $\kappa_\varepsilon(x, x')$, während $k(x)$ immer noch der Kovarianzvektor des unverrauschten Prozesses aufgrund der Unkorreliertheit von $Z(x)$ und $\varepsilon(x')$ ist:

$$\begin{aligned} Cov[Z(x), Z_\varepsilon(x')] &= Cov[Z(x), Z(x') + \varepsilon(x')] \\ &= Cov[Z(x), Z(x')] + Cov[Z(x), \varepsilon(x')] \\ &= Cov[Z(x), Z(x')] = \kappa(x, x'). \end{aligned} \tag{33}$$

Damit ist $k(x_i)$ kein Vektor der Matrix $K_\varepsilon = [\kappa_\varepsilon(x_i, x_j)]_{1 \le i,j \le n}$, was direkt dazu führt, dass $\widehat{Z}(x_i) \neq z(x_i)$ und $\widehat{s}^2(x_i) > 0$ für $i = 1, \ldots, n$.

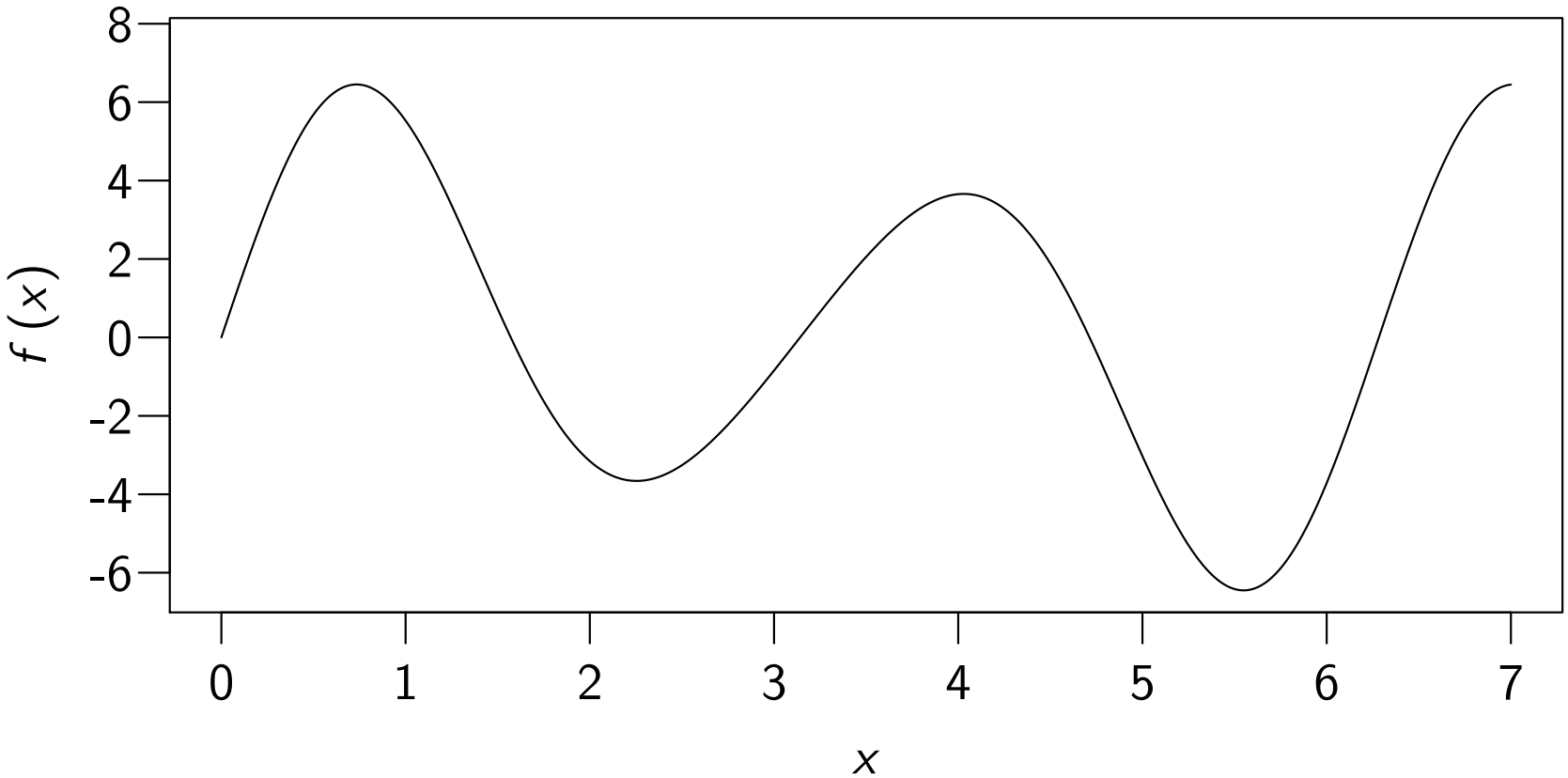
9.3 Modellbasierte Optimierung: Idee

Sei $f : \mathbb{D} \mapsto \mathbb{R}$ eine Blackbox-Funktion, also eine Funktion

- über deren Eigenschaften nichts oder nur wenig bekannt ist und
- deren Auswertung teuer im Sinne von z.B. tatsächlich anfallenden Kosten und/oder zeitlichen Aufwand ist.

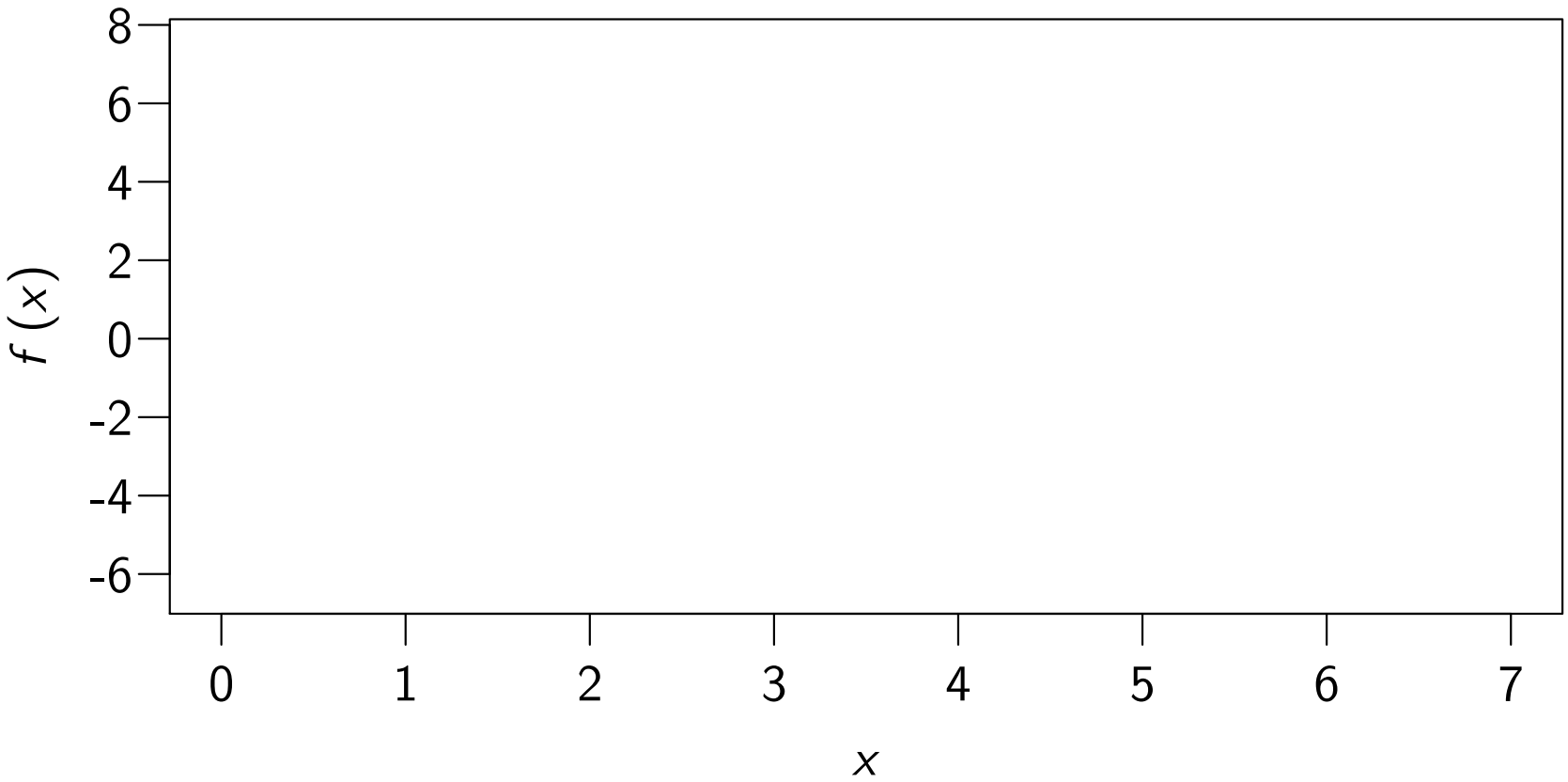
Um das Optimum bzw. die optimalen Parameter $x \in \mathbb{D} \subseteq \mathbb{R}^d$ ($d \in \mathbb{N}$) zu ermitteln, wird bei der modellbasierten Optimierung der unbekannte Funktionsverlauf von f modelliert und mit Hilfe des weit weniger teuren *Surrogatmodells*, z.B. Kriging, das Optimum gesucht.

9.3 Modellbasierte Optimierung: Idee



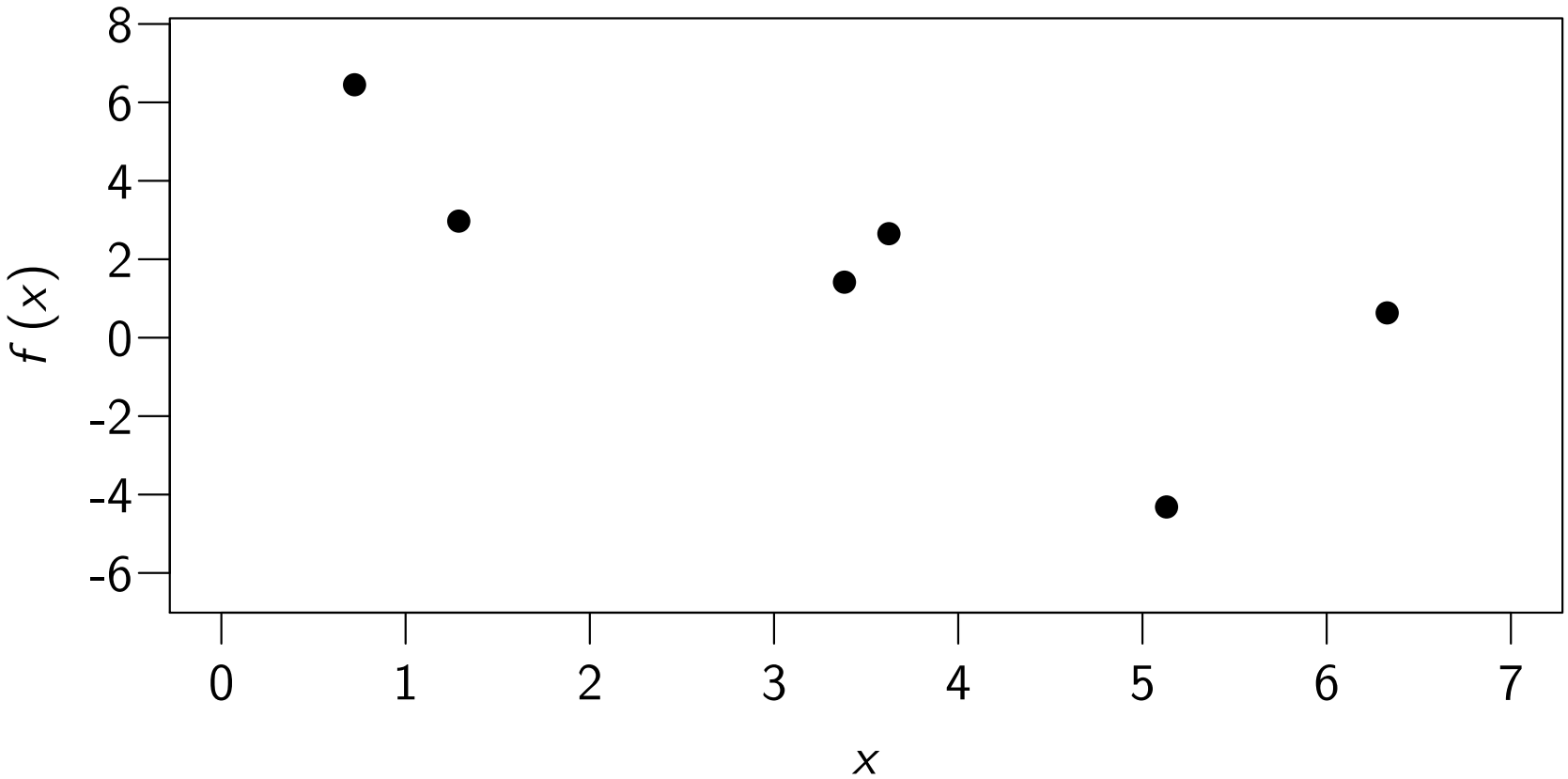
Sei $f : \mathbb{D} \mapsto \mathbb{R}$ Funktion, deren Minimum wir finden wollen.

9.3 Modellbasierte Optimierung: Idee



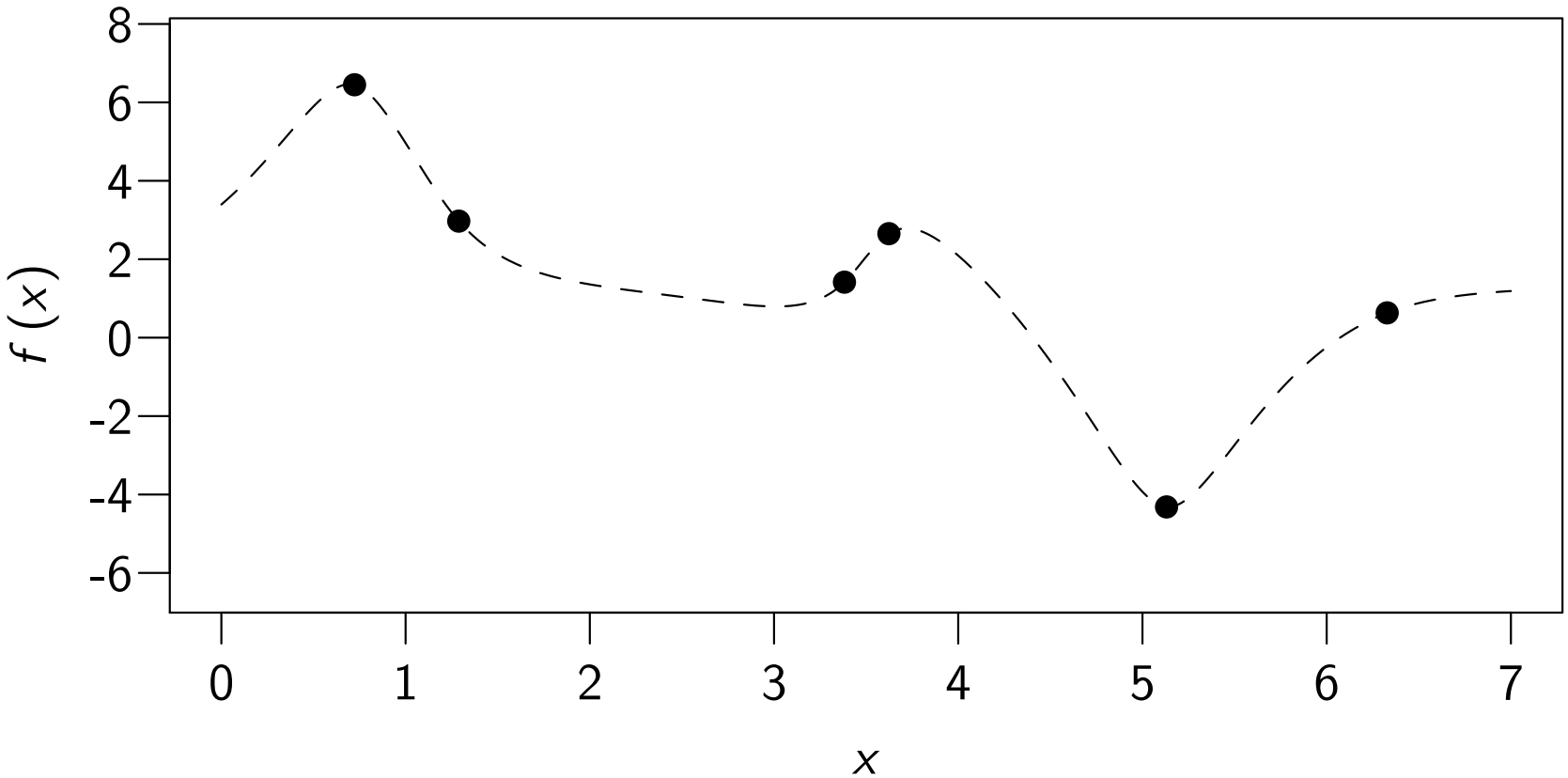
Sei $f : \mathbb{D} \mapsto \mathbb{R}$ Funktion, die wir aber nicht kennen (Black-Box).

9.3 Modellbasierte Optimierung: Idee



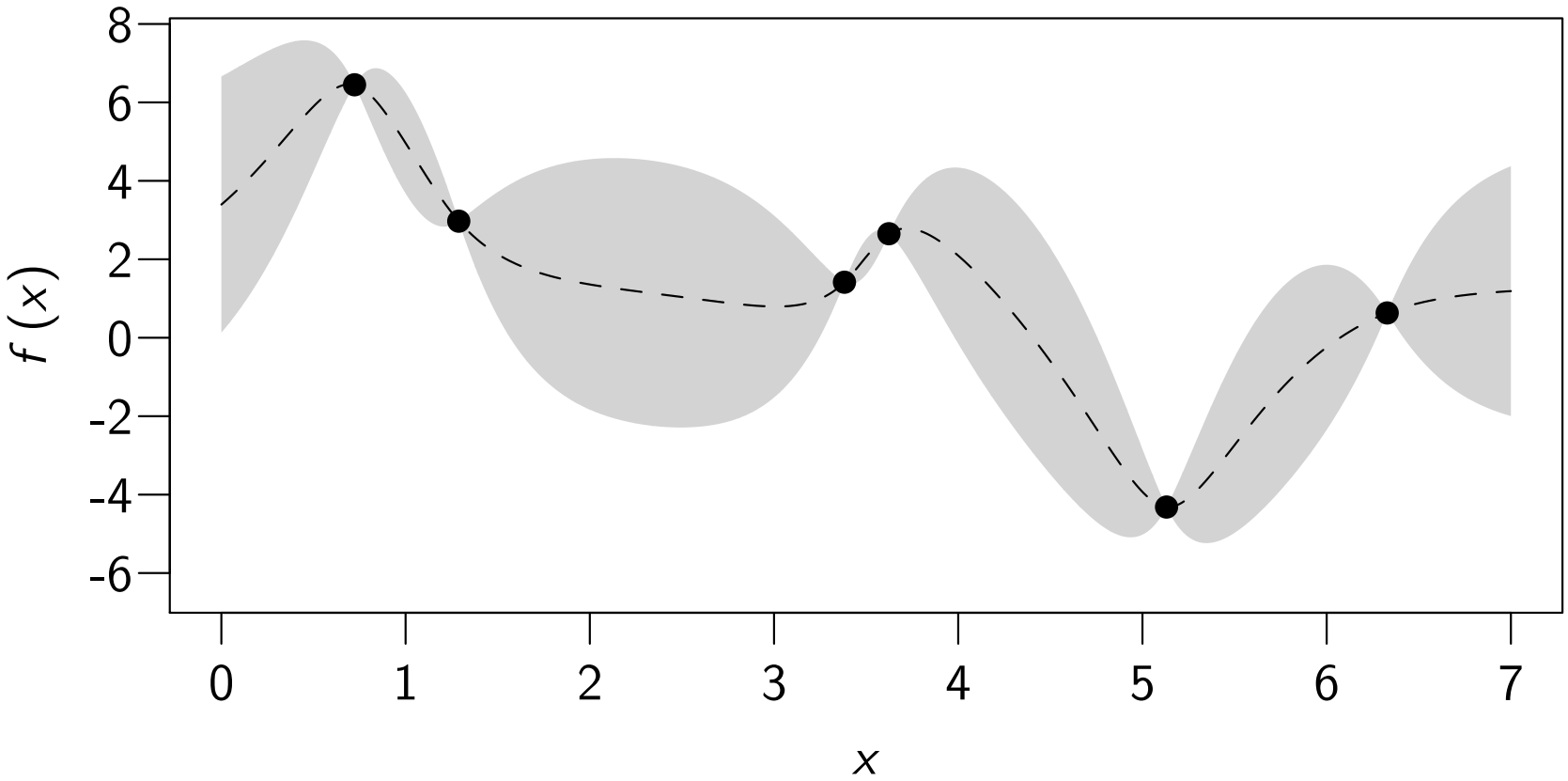
Werte f in $x_1, \ldots, x_n$ (hier $n = 6$) eines raumfüllenden Designs aus: $f(x_i) = z_i$

9.3 Modellbasierte Optimierung: Idee



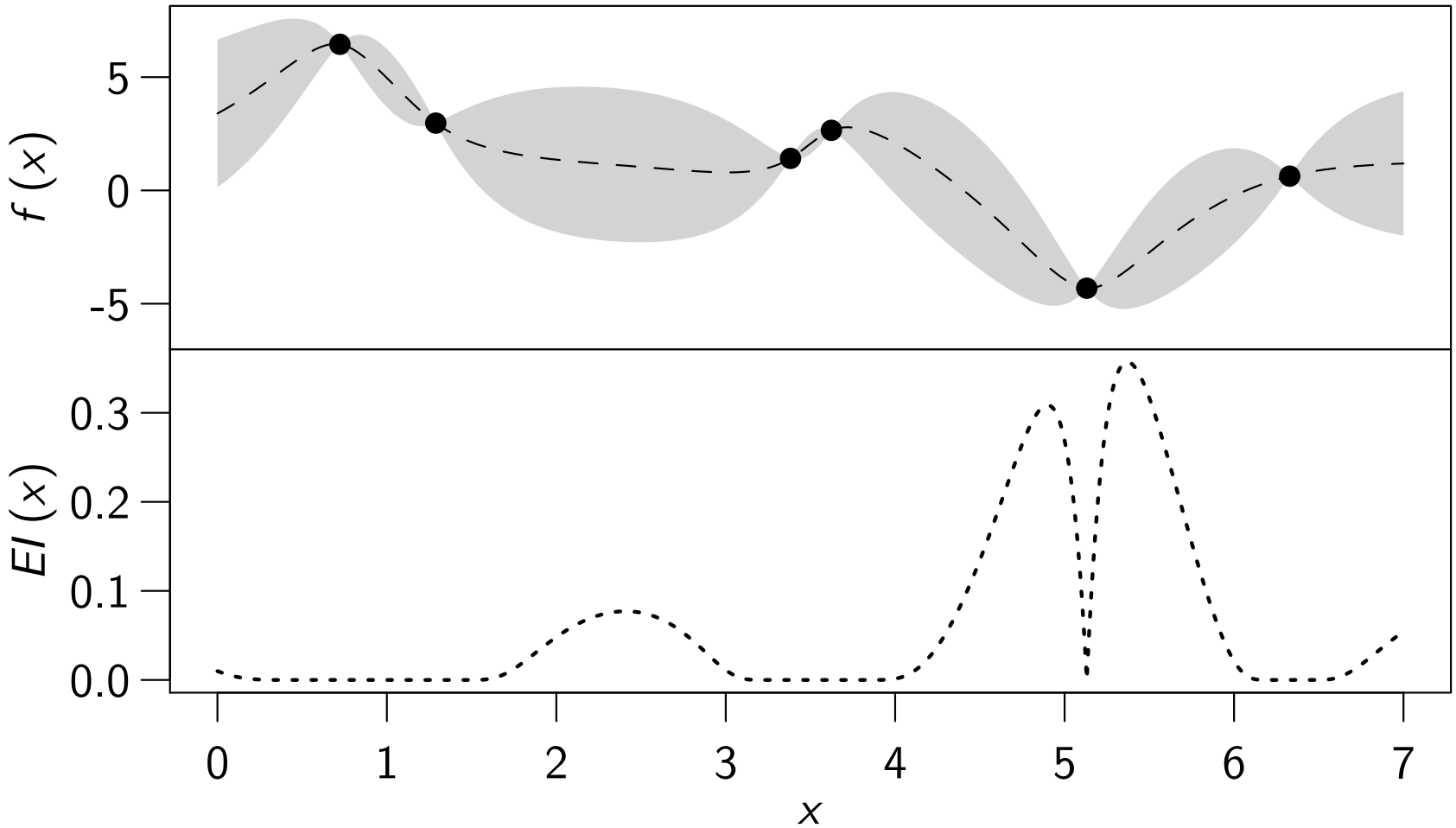
Bestimme den Kriging-Mittelwert gegeben $z = (z_1, \ldots, z_n)^T$: $\widehat{Z}(x)$

9.3 Modellbasierte Optimierung: Idee



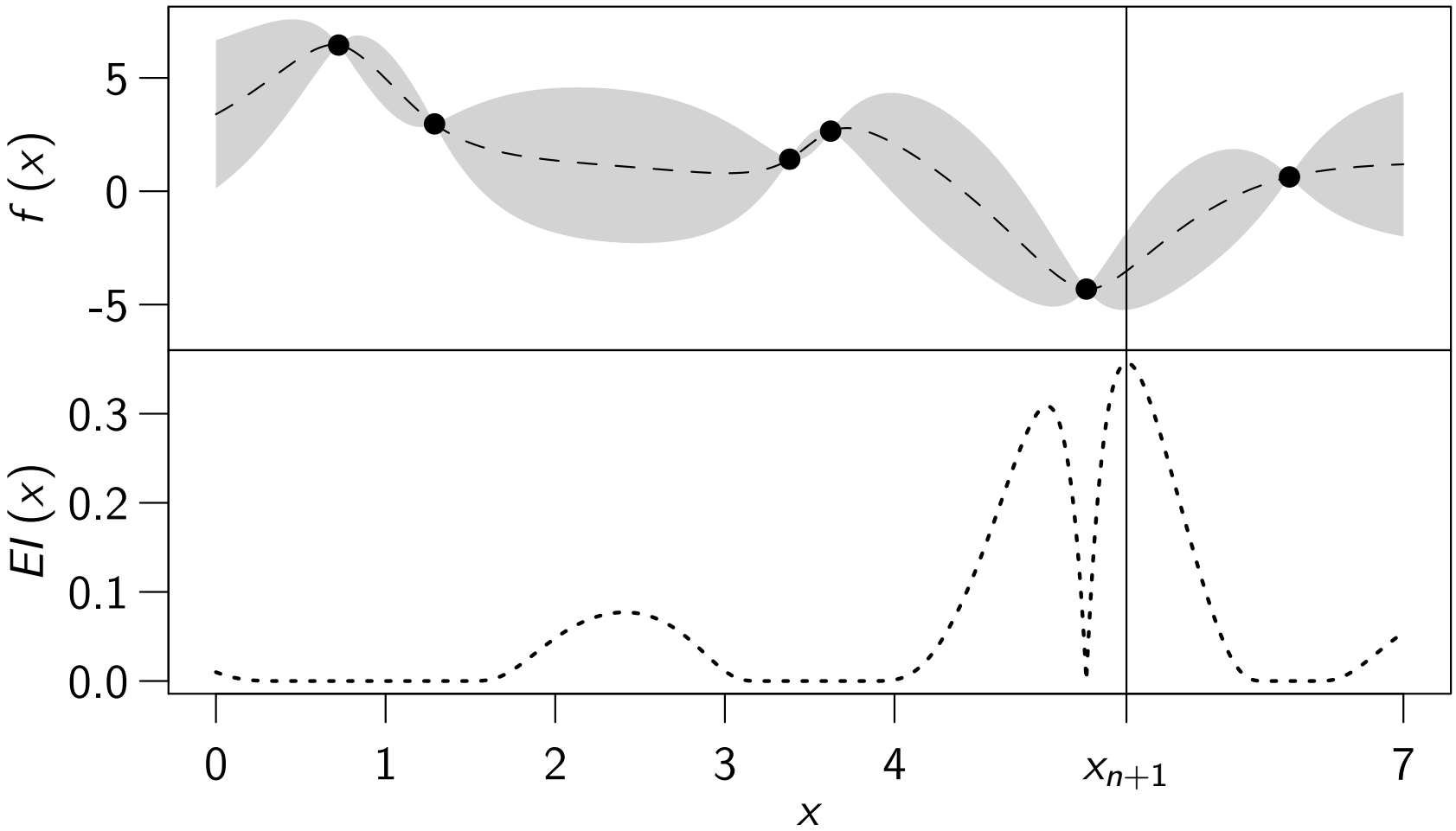
Bestimme den Kriging-Varianz gegeben $z = (z_1, \ldots, z_n)^T$: $\widehat{s}^{\,2}(x)$

9.3 Modellbasierte Optimierung: Idee



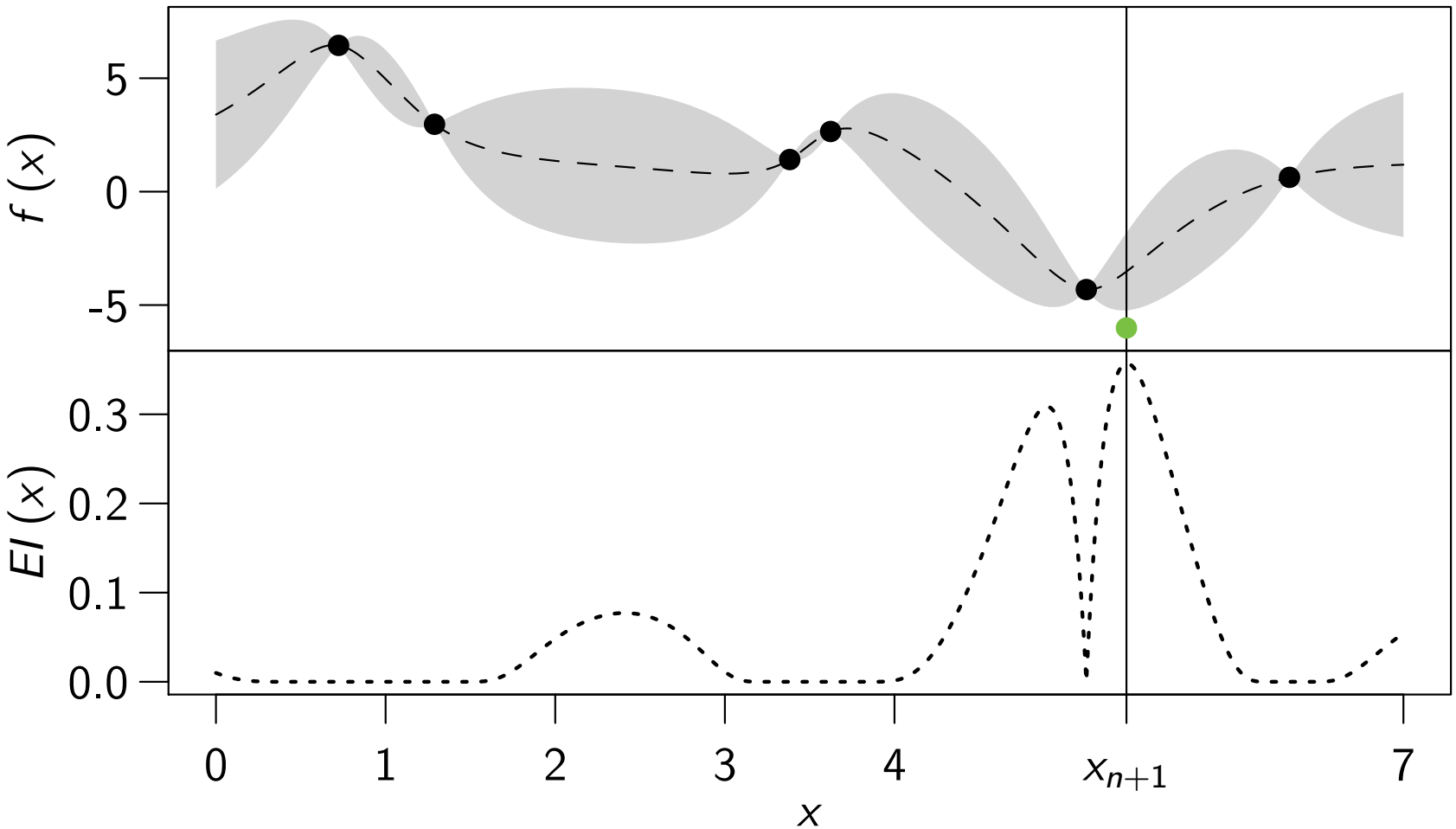
Berechne Infill-Kriterium (hier: Expected Improvement), um den nächsten Auswertungspunkt zu bestimmen.

9.3 Modellbasierte Optimierung: Idee



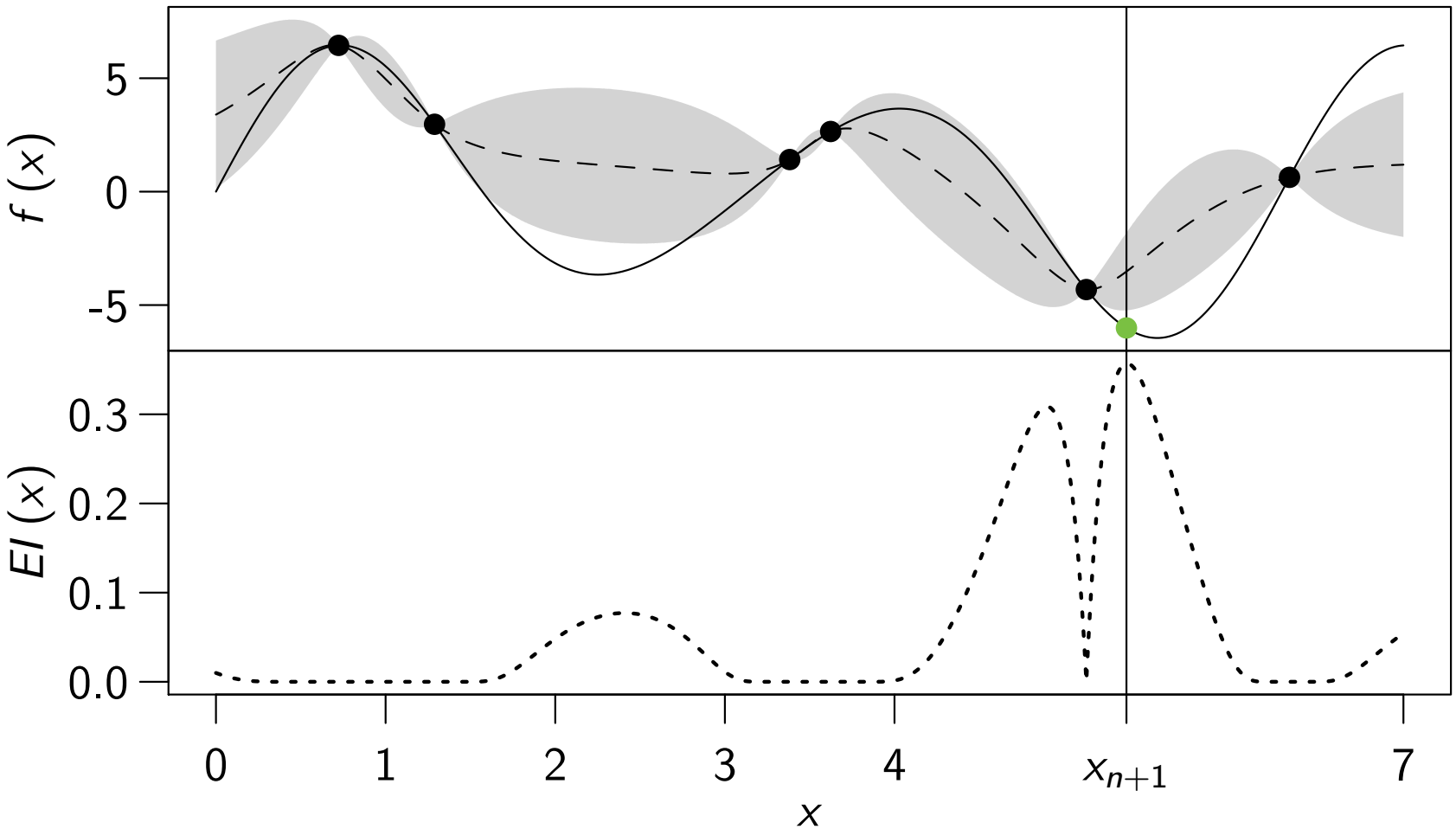
Nächster Punkt für Auswertung: $x_{n+1} = \arg\max_{x \in \mathbb{D}} EI(x)$

9.3 Modellbasierte Optimierung: Idee



Werte nächster Punkt x_{n+1} aus: $z_{n+1} = f(x_{n+1})$

9.3 Modellbasierte Optimierung: Idee



Werte nächster Punkt x_{n+1} aus: $z_{n+1} = f(x_{n+1})$

9.3 Modellbasierte Optimierung: Skizze

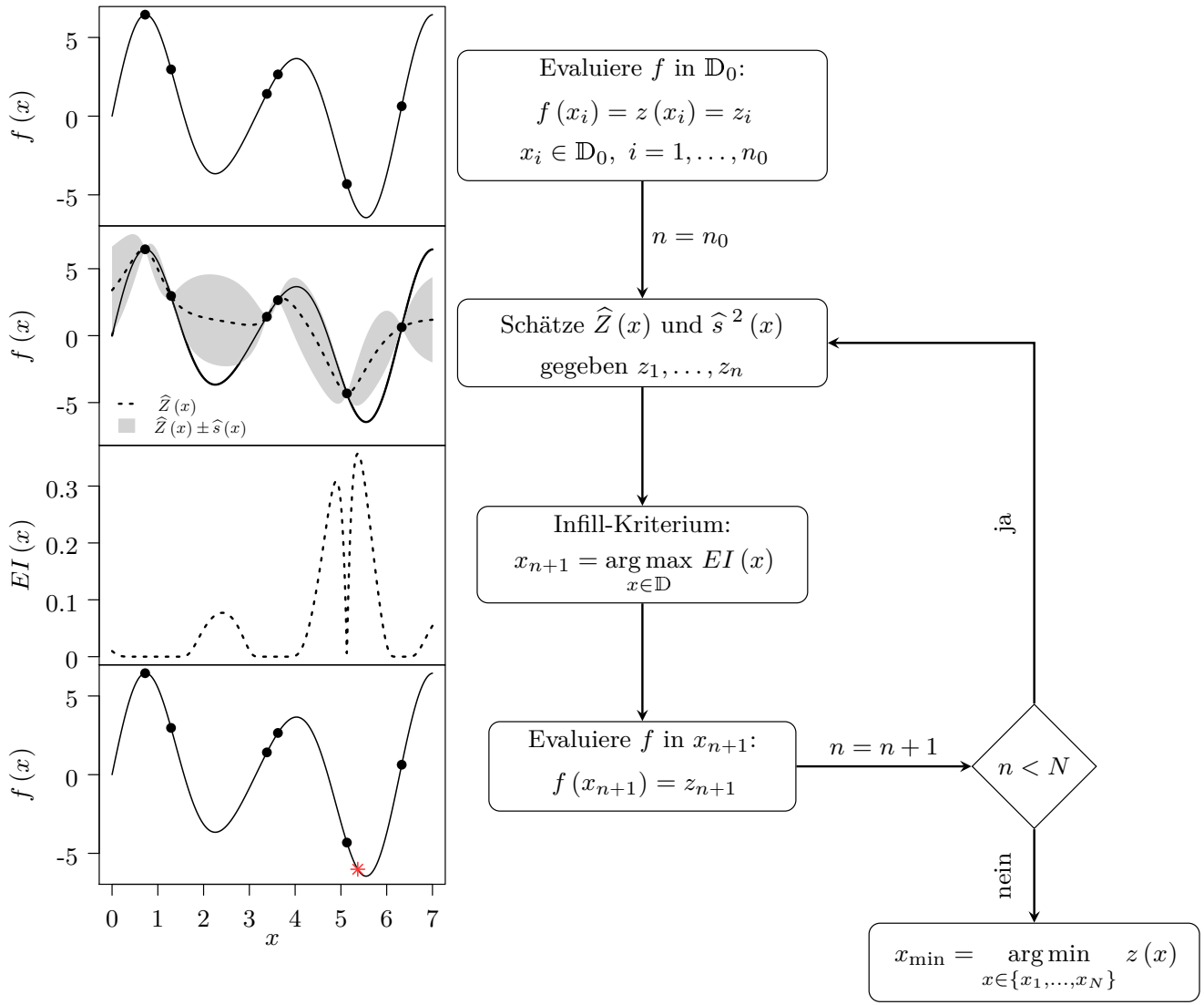


Abb. 9.4 : Beispiel für den Verlauf einer modellbasierten Optimierung (Modell: Kriging, Infill Kriterium: Expected Improvement (*EI*), Abbruchkriterium: Anzahl Iterationen).

9.3 Modellbasierte Optimierung: Schritte

Das Vorgehen gliedert sich dabei in fünf Teilschritte (vgl. Abbildung 4):

1. Die Funktion f wird in den Punkten eines üblicherweise raumfüllenden Versuchsplans $\mathbb{D}_0 \subset \mathbb{D}$ ausgewertet.
2. Die Surrogatmodellparameter werden geschätzt.
3. Suche nach dem nächsten Auswertungspunkt.
4. Evaluiere die Zielfunktion.
5. Falls Abbruch, gib den bisher besten Auswertungspunkt zurück oder starte erneut bei 2.

9.3 Modellbasierte Optimierung: Schritte

Nun im Detail zu den Schritten 1–3:

1. Die Funktion f wird in den Punkten eines üblicherweise raumfüllenden Versuchsplans $\mathbb{D}_0 \subset \mathbb{D}$ ausgewertet.
   - Damit wird eine Datengrundlage für die Schätzung der Surrogatmodellparameter in Schritt 2 geschaffen.
   - Mögliches raumfüllendes Design z.B. vollfaktorieller Plan oder Latin-Hypercube [106].
   - Die Anzahl n_0 der Punkte im Initialdesign orientiert sich oft an der Dimension d des Parameterraumes. [81] empfehlen $n_0 = 10d$ Initialauswertungen.

   Nach Auswertung des Initialdesigns startet der Optimierungsprozess. Die Iteration wird beendet, wenn ein Abbruchkriterium (z.B. das Erreichen einer vorher festgelegten Anzahl N an Iterationsschritten) erfüllt ist.

9.3 Modellbasierte Optimierung: Schritte

2. Die Surrogatmodellparameter werden geschätzt.
   - Wahl des Modells ist abhängig von den Eigenschaften der Funktion f und von der Beschaffenheit des Parameterraumes $\mathbb{D}$. Üblicherweise wird bei deterministischen Computerexperimenten mit reellwertigen Parametern $x \in \mathbb{D} \subseteq \mathbb{R}^d$ das gewöhnliche Kriging-Modell $Z(x)|Z$ verwendet.
     In seiner interpolierenden Art hat es bei deterministischen Anwendungen den Vorteil, dass es keine Unsicherheit bezüglich bereits ausgewerteter Punkte lässt.
   - Zahlreiche andere Möglichkeiten interpolierende Modelle zu konstruieren, z.B. Splines.
   - Meist wird hierbei eine Kombination aus der gewichteten Summe von Polynomen und der gewichteten Summe von bestimmen Basisfunktionen, welche um bereits beobachtete Punkte zentriert sind (z.B. dem Euklidischen Abstand), verwendet.
   - Eine ausführliche Diskussion unterschiedlicher Modelle ist bei [82] zu finden.

9.3 Modellbasierte Optimierung: verrauschte Beobachtungen

Ist die zu optimierende Zielfunktion f stochastisch, in dem Sinne, dass die Werte der Funktion verrauscht sind, so bietet sich das Kriging-Modell mit Nugget-Effekt $Z_\varepsilon(x)|Z_\varepsilon$ oder das noisy Kriging-Modell $Z(x)|Z_\varepsilon$ an.

- Die folgende Abbildung zeigt die Vorhersage $\widehat{Z}(x)$ und die Unsicherheit ausgedrückt als die ausgefüllte Fläche zwischen $\widehat{Z}(x) \pm \widehat{s}(x)$.
- Der Verlauf der Vorhersage mit $Z_\varepsilon(x)|Z_\varepsilon$ und der Vorhersage mit $Z(x)|Z_\varepsilon$ unterscheiden sich in der Hinsicht, dass die Vorhersage mit $Z_\varepsilon(x)|Z_\varepsilon$ (blaue Kurve) Ausschläge zu den Beobachtungen aufweist und im Fall von nur einer Beobachtung in x diese auch interpoliert, was zu einem unstetigen Verlauf führt.
- Das R-Paket `mlrMBO` [9] bietet hier die Option `jitter`, die bei der Berechnung der Vorhersage eines bereits ausgewerteten Punktes x_i $(i = 1, \ldots, n)$ statt $\widehat{Z}(x_i)$ eine benachbarten Wert $\widehat{Z}(x_i + c)$ zurückgibt, wobei c eine Größenordnung von 10^{-12} hat. Das Ergebnis ist die gelb gestrichelte Kurve der Vorhersage.

9.3 Modellbasierte Optimierung: verrauschte Beobachtungen

- Ein weiterer Unterschied zwischen den beiden Modellen besteht bei der geschätzten Unsicherheit der Vorhersage. Im Vergleich zum Modell mit Nugget-Effekt $(Z_\varepsilon(x)|Z_\varepsilon)$ ist beim noisy Kriging-Modell $(Z(x)|Z_\varepsilon)$ die Unsicherheit etwas kleiner und fällt auch bei bereits beobachteten Punkten nicht auf Null (vgl. rote Fläche).
- Legende für folgende Grafik:
  - schwarze Linie: wahre Funktion;
  - schwarze Punkte: verrauschten Beobachtungen;
  - blaue Fläche: $\widehat{Z}(x) \pm \widehat{s}(x)$ (Kriging mit Nugget-Effekt) $Z_\varepsilon(x)|Z_\varepsilon$;
  - blau bzw. gelb (`jitter`): zugehörige Vorhersage;
  - rote Fläche: $\widehat{Z}(x) \pm \widehat{s}(x)$ (noisy Kriging) $Z(x)|Z_\varepsilon$;
  - rot: zugehörige Vorhersage.

9.3 Modellbasierte Optimierung: verrauschte Beobachtungen

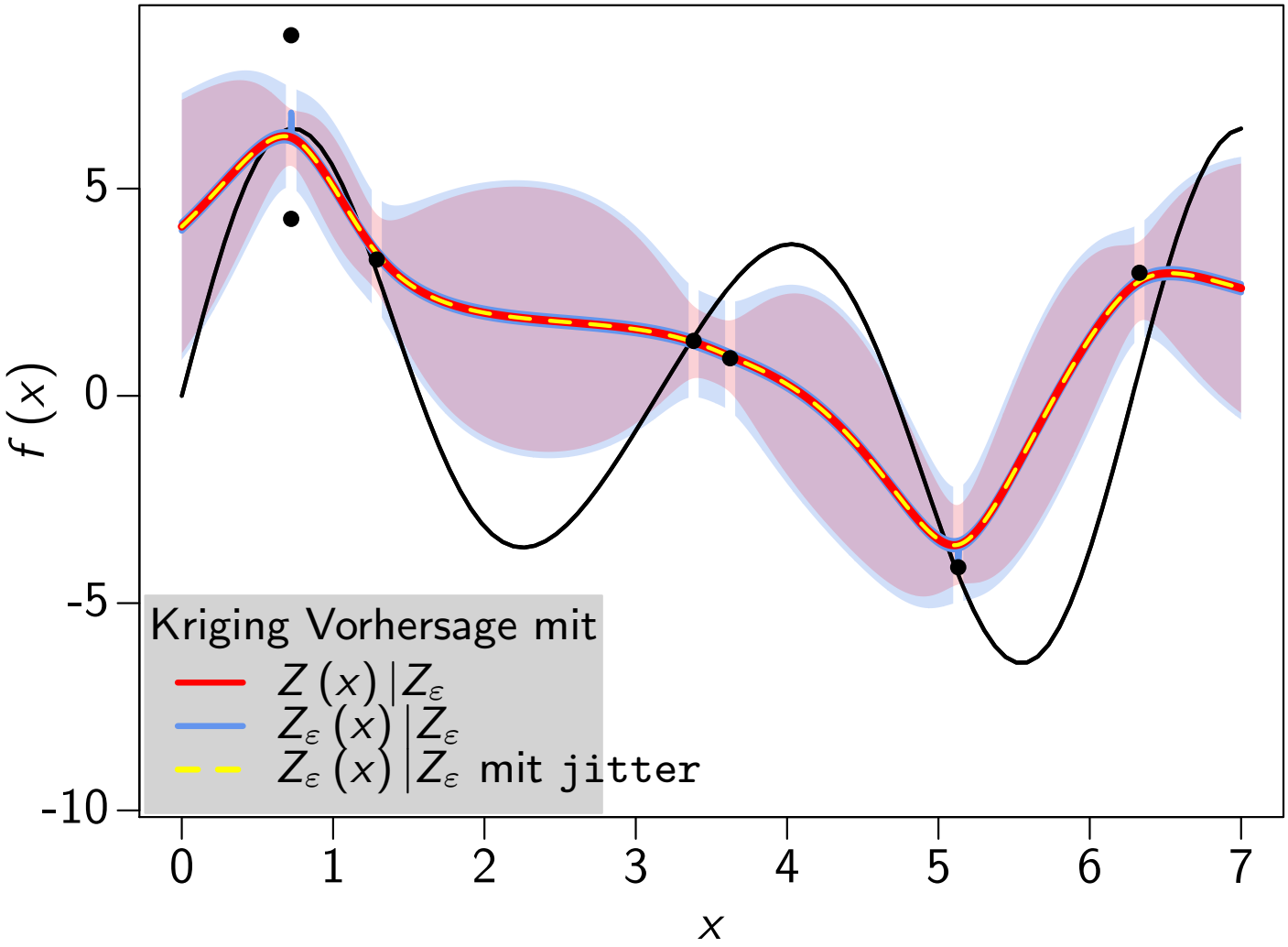


Abb. 9.5 : ausführliche Legende siehe vorige Folie.

9.3 Modellbasierte Optimierung: Schritte

3. Suche nach dem nächsten Auswertungspunkt.
   - Da nun Verlauf der Modellfunktion bekannt ist, kann nächster Punkt mit „Bedacht“ gewählt werden:
   - Suchstrategie als ausgewogene Kombination von Exploitation und Exploration,
   - Optimum einerseits möglichst genau bestimmen und andererseits Chance erhöhen wirklich das globale und nicht nur ein lokales Optimum zu finden.
   - Bei Exploitation kann sich die Punktwahl z.B. am Minimum der Modellvorhersage orientieren.
     Stimmt das Minimum der Modellvorhersage mit einem bereits ausgewerteten Punkt überein, käme die Optimierung zum Stillstand. In [82] werden Alternativen diskutiert.
   - Bei Exploration bietet das Kriging einen entscheidenden Vorteil, da hierbei neben der Vorhersage $\widehat{Z}(x)$ auch die Unsicherheit dieser Vorhersage $\widehat{s}^{\,2}(x)$ geschätzt wird. Das Abstandsmaß in der Kovarianzfunktion bewirkt dabei, dass $\widehat{s}^{\,2}(x)$ größere Werte annimmt, je weiter ein Punkt x von bereits beobachteten Punkten $x_1, \ldots, x_n$ entfernt ist. Auf diese Weise lassen sich die unbekannten Regionen im Parameterraum leicht identifizieren.

9.3 Modellbasierte Optimierung: Kriterien

Für die Kombination von Exploitation und Exploration wurden im Laufe der Zeit einige Kriterien (auch Infill-Kriterien genannt) vorgeschlagen;

- Das *Lower-Confidence-Bound*

$$LCB(x) = \widehat{Z}(x) - c\widehat{s}(x) \tag{34}$$

mit $c > 0$ von [24] ist in seiner Form sehr intuitiv.

- In der zweiten Grafik von oben in Abbildung 4 markiert die untere Grenze der grauen Fläche gerade das Lower-Confidence-Bound mit $c = 1$.
- Als nächster Punkt wird vorgeschlagen:

$$x_{n+1} = \arg\min_{x \in \mathbb{D}} LCB(x). \tag{35}$$

- Die Wahl des Parameters c ist hierbei entscheidend. Betrachten wir z.B. Abbildung 4 mit $c = 1$, so wird $x_{n+1} = 5.346$ gewählt. Mit $c = 5$ erhalten wir $x_{n+1} = 2.354$, da in diesem Bereich die Unsicherheit besonders groß ist.

9.3 Modellbasierte Optimierung: Kriterien

- Das *Expected Improvement* hängt von keinen weiteren Parametern ab findet wohl am häufigsten Verwendung:

$$EI(x) = \left(z_{\min} - \widehat{Z}(x)\right) \Phi\left(\frac{z_{\min} - \widehat{Z}(x)}{\widehat{s}(x)}\right) + \widehat{s}(x)\,\varphi\left(\frac{z_{\min} - \widehat{Z}(x)}{\widehat{s}(x)}\right) \quad (36)$$

für $\widehat{s}^{\,2}(x) > 0$ und 0, sonst. Dabei ist $z_{\min} = \min_{1 \le i \le n} z(x_i)$ das nach n ausgewerteten Punkten $x_1, \ldots, x_n$ realisierte Minimum der Funktion f. Φ und φ bezeichnen die Verteilungsfunktion und die Dichte der Standardnormalverteilung.

  - Das Konzept fand bereits 1978 Anwendung [113] und wurde später [81] im Zusammenhang mit effizienter globaler Optimierung (EGO) ein Standard.
  - Durch Maximieren der Funktion erhalten wir als nächsten Punkt

$$x_{n+1} = \arg\max_{x \in \mathbb{D}} EI(x). \quad (37)$$

9.3 Modellbasierte Optimierung: Kriterien

- Im Fall einer stochastischen zu optimierenden Zielfunktion f ist das Minimum $z_{\min}$ nicht exakt bekannt, da wir annehmen, dass alle Beobachtungen $z(x_1) = f(x_1), \ldots, z(x_n) = f(x_n)$ Realisationen des verrauschten Prozesses $Z_\varepsilon(x)$ sind und daher mit einer Fehlervarianz σ_ε^2 um den wahren Funktionswert schwanken.
- Vorgeschlagen wurde [75] statt $z_{\min} = \min_{1 \leq i \leq n} z(x_i)$ die Vorhersage $\widehat{Z}(x^\star)$ an der effektiv besten Stelle

$$x^\star = \underset{x \in \{x_1, \ldots, x_n\}}{\arg\min} \; \widehat{Z}(x) + c\widehat{s}(x) \tag{38}$$

  zu wählen, wobei die Autoren $c = 1$ empfehlen.
- Statt des ursprünglichen Expected Improvements $EI_{z_{\min}}(x)$ wird nun $EI_{\widehat{Z}(x^\star)}(x)$ verwendet.
- Zusätzlich wird die Fehlervarianz σ_ε^2 durch einen Strafterm berücksichtigt, so dass das Augmented Expected Improvement für $\widehat{s}^{\,2}(x) > 0$ gegeben ist als

$$AEI(x) = EI_{\widehat{Z}(x^\star)}(x) \left[1 - \sqrt{\frac{\sigma_\varepsilon^2}{\widehat{s}^{\,2}(x) + \sigma_\varepsilon^2}} \right]. \tag{39}$$

9.3 Modellbasierte Optimierung: Augmented Expected Improvement

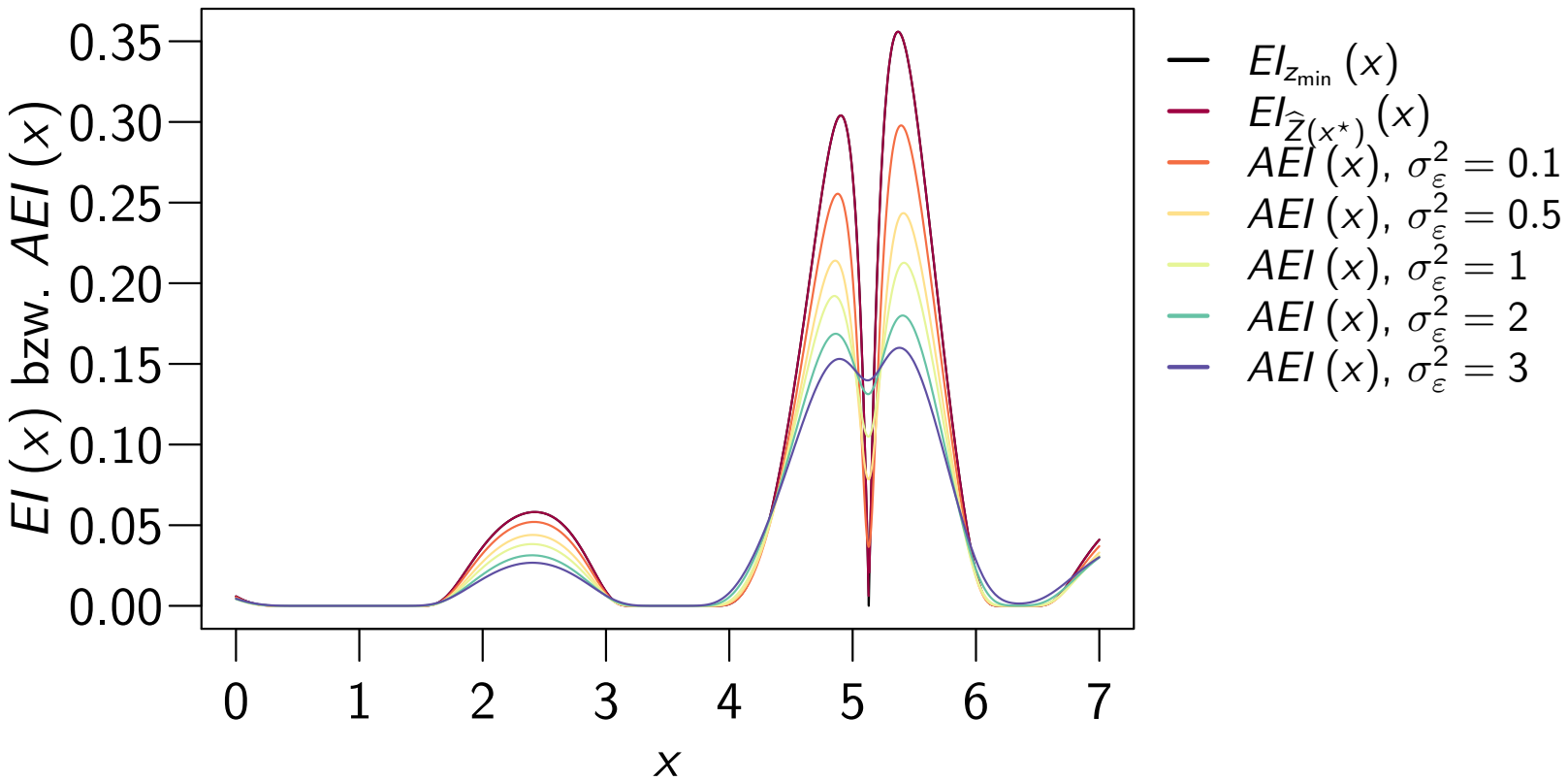


Abb. 9.6 : Beispiele für (Augmented) Expected Improvement.

9.3 Modellbasierte Optimierung: Software

Um in der Praxis die Stelle des Optimums x_{n+1} für ein gegebenes Infill-Kriterium zu finden, bietet das R-Paket `mlrMBO` [9] neben dem evolutionären Algorithmus u.a. auch den sogenannten Focus Search-Algorithmus [72] an.

Beim Focus Search wird der Parameterraum sukzessiv um das aktuelle Optimum verkleinert, wobei die Prozedur mehrmals wieder von vorne anfängt, was für eine ausgewogene Mischung aus Exploitation und Exploration sorgt. Der folgenden Pseudocode zeigt das genaue Vorgehen beim Minimieren eines Infill-Kriteriums c. Soll das Kriterium maximiert werden, wie dies beim Expected Improvement der Fall ist, so wird $-c$ minimiert.

9.3 Modellbasierte Optimierung: Algorithmus

Input : infill criterion $c : \mathbb{D} \to \mathbb{R}$
control parameters $n_{\texttt{restart}}$, $n_{\texttt{iters}}$, $n_{\texttt{points}}$

for $u \in \{1, \ldots, n_{\texttt{restart}}\}$ **do**
  Set $\tilde{\mathbb{D}} = \mathbb{D}$
  for $v \in \{1, \ldots, n_{\texttt{iters}}\}$ **do**
    generate random design $\mathcal{D} \subset \tilde{\mathbb{D}}$ of size $n_{\texttt{points}}$
    compute $x^*_{u,v} = (x^*_1, \ldots, x^*_d)^T = \arg\min_{x \in \mathcal{D}} c(x)$
    # shrink $\tilde{\mathbb{D}}$ by focusing on $x^*_{u,v}$:
    for each space dimension $j \in \{1, \ldots, d\}$ **do**
      if $\tilde{\mathbb{D}}_j$ numeric with values $[l_j, u_j]$ **then**
        $l_j = \max\{l_j, x^*_j - \frac{1}{4}(u_j - l_j)\}$
        $u_j = \min\{u_j, x^*_j + \frac{1}{4}(u_j - l_j)\}$
        $\tilde{\mathbb{D}}_j = [l_j, u_j]$
      end
      if $\tilde{\mathbb{D}}_j$ categorical with values $\{x_{j1}, \ldots, x_{js}\}$, $s > 2$ **then**
        sample one category $\bar{x}_j$ uniformly from $\tilde{\mathbb{D}}_j \setminus \{x^*_j\}$
        $\tilde{\mathbb{D}}_j = \tilde{\mathbb{D}}_j \setminus \{\bar{x}_j\}$
      end
    end
  end
end
Result : $x^* = \underset{\substack{u \in \{1,\ldots,n_{\texttt{restart}}\} \\ v \in \{1,\ldots,n_{\texttt{iters}}\}}}{\arg\min} \; c(x^*_{u,v})$

Abb. 9.7 : Optimierung Infill-Kriterium: Focus Search [72].